EPFL

Thèse n° 7808

Deterministic and Statistical Approaches to Quantum Chemistry

Présentée le 4 septembre 2020

à la Faculté des sciences de base
Laboratoire de design moléculaire computationnel
Programme doctoral en chimie et génie chimique

pour l'obtention du grade de Docteur ès Sciences

par

Alberto FABRIZIO

Acceptée sur proposition du jury

Prof. B. Fierz, président du jury
Prof. A.-C. Corminboeuf, directrice de thèse
Dr S. Steinmann, rapporteur
Dr P.-F. Loos, rapporteur
Prof. J. Vanicek, rapporteur

École
polytechnique
fédérale
de Lausanne

2020

> This turkey found that he was fed at 9 a.m. However, being a good inductivist, he did not jump to conclusions. He waited until he had collected a large number of observations of the fact that he was fed at 9 a.m. Finally, his inductivist conscience was satisfied and he carried out an inductive inference to conclude, "I am always fed at 9 a.m.". Alas, this conclusion was shown to be false in no uncertain manner when, on Christmas eve, instead of being fed, he had his throat cut.
>
> — Bertrand Russel

This thesis is dedicated to my wife and to my family.

Acknowledgements

All my gratitude goes to Prof. Clémence Corminboeuf for her constant support, guidance and her boundless ability to inspire ambition and perseverance. In these last four years, she contributed not only to improve my scientific work, but also myself as a person. I am particularly grateful to her for never preventing me to follow my curiosity and for challenging me to achieve goals that I thought well beyond my possibilities.

I am grateful also to all the members of LCMD, past and present, who never failed to inspire and support me and my research. A special thanks goes to former lab members Dr. Riccardo Petraglia, Dr. Antonio Prlj and Dr. Benjamin Meyer, whose friendship has been determinant in the hardest moments of the PhD. Dr. Matthew Wodrich is also warmly acknowledged not only for promoting the development of my writing and presenting skills, but also for co-founding the 10K Tuesday tradition (which has been crucial for my physical and mental health). In addition, (DJ) Raimon Fabregat, Kun-Han Lin, Boodsarin Sawatlon and Veronika Juraskova deserve a special mention, as they shared with me the joy and the pain of doctoral studies for years. To all of them, I wish a successful thesis and a bright career. I was lucky enough to meet a new generation of highly skilled PhD students: Terry Blaskovits, Simone Gallarati and Ksenia Briling. Their passion has rekindled my love for quantum chemistry. I really hope that Dr. Sergio (Sergi) Vela and Dr. Maria Fumanal will be able to accomplish their projects both in life and research. I wish to both the best of luck. A special thanks goes to Véronique Bujard and Dr. Daniel Jana, truly pillars of LCMD. I am happy to welcome two new lab members Dr. Shubhajit Das and Dr. Marc Garner. I am really looking forward to working with you.

Nothing in work or life could be accomplished without the support of a loving and caring family and affectionate friends. I am particularly grateful to my parents, whose example has always inspired me to reach the highest standards. I could not be a better man without my lovely wife, whose constant support and seemingly infinite patience are truly an everyday gift. To my sister, I really hope your dreams and ambitions would become true. My wife's family also deserves a special mention for the support I received: I felt part of the clan since the first day we met.

I am also extremely grateful to prof. Beat Fierz, prof. Jiri Vanicek, Dr. Stephan Steinmann and Dr. Pierre-François Loos for having accepted to be on my jury. Finally, the National Centre of Competence in Research (NCCR) "Materials' Revolution: Computational Design and Discovery of Novel Materials (MARVEL)" of the Swiss National Science Foundation (SNSF) and the EPFL are acknowledged for financial support.

Lausanne, 12 Mars 2020 A. F.

Abstract

The field of quantum chemistry has recently undergone a series of paradigm shifts, including a boom in machine learning applications that target the electronic structure problem. Along with these technological innovations, the community continues to identify shortcomings in traditional KS-DFT approaches and develop improved approximations. The original work presented in this thesis addresses a selection of open questions along these two lines. Specifically, the thesis is structured to reflect the ongoing advancement of traditional (*deterministic*) approaches toward more recent examples exploiting (*statistical*) non-linear regression techniques.

The first section of the thesis focuses on analyzing the performance of approximate density functionals and dispersion correction schemes on chemical situations that are not well-represented in standard benchmark databases of van der Waals complexes. A first example discusses how the synergy between delocalization error and London dispersion interactions in asymmetrically charged radical cation dimers remains problematic, even for the most recent density functionals. Solutions are provided to improve the description of these systems that are typical charge-carrier in organic electronic materials. While this first chapter focuses on non-covalent interactions between molecules in their electronic ground-state, very little is known about the consequences of an incomplete treatment of London dispersion interactions involving molecules upon photo-excitation. Using the prototypical stilbene photoswitch as a working example, the second chapter demonstrates that completely neglecting these interactions in the excited states leads to qualitative failures in the description of the photodeactivation process. The conclusions presented in this chapter apply broadly to any photoswitch functionalized with large and polarizable side chains.

In contrast to traditional (*deterministic*) quantum chemistry, machine learning-based variants are still in their infancy when facing the challenge of targeting fundamental, albeit complex, quantum chemical objects. The subsequent chapters describe the development and application of machine-learning techniques to predict the molecular electron density [$\rho(\boldsymbol{r})$] using an atom-centered representation compatible with symmetry-adapted Gaussian process regression (SA-GPR). Concrete applications of the framework are shown for a chemically rich set of dimers, whose predicted electron densities serve to compute covalent and non-covalent interaction fingerprints, electrostatic potentials, as well as quantitative interaction energies. The transferability of the model is demonstrated by the accurate prediction of $\rho(\boldsymbol{r})$ for a set of pentapeptides. Combining transferability and accuracy, our regression framework grants access to the density information of complex chemical systems at a fraction of the traditional *ab-initio* computational cost.

The final chapter exploits the complementarity of both strategies and proposes a machine learning framework capable of quantifying the deviation of approximate density functionals from the piecewise linearity condition of exact DFT. The predicted curvature information is applied both for restoring the correspondence between the Kohn-Sham HOMO eigenvalue and the first ionization potential in optimally tuned DFT functionals as well as to provide a large-scale analysis of the relationship between the deviation from the piecewise-linearity condition and the chemical and structural patterns.

Overall, the work discussed in this thesis is part of a more comprehensive effort to extend the applicability of KS-DFT to uncommon chemical situations and to increasingly complex molecular systems by leveraging the latest advances in "quantum machine-learning".

Keywords: Quantum Chemistry, Density Functional Theory, Machine Learning, Electron Density, Delocalization Error, Non-Covalent Interactions

Résumé

La chimie quantique a récemment subi une série de changements de paradigme, notamment un boom des applications d'apprentissage automatique qui ciblent les problèmes de structure électronique. Parallèlement à ces innovations technologiques, la communauté scientifique continue d'identifier les limites des approches KS-DFT traditionnelles et de développer de meilleures approximations. Le travail original présenté dans cette thèse aborde une sélection de questions ouvertes dans ces deux domaines. Plus précisément, la thèse est structurée de manière à refléter l'avancement des méthodes traditionnelles (*déterministes*) vers approches plus récentes, qui utilisent des techniques de régression non linéaire (*statistique*).

La première section de la thèse analyse la performance des fonctionnelles de la densité et des corrections de dispersion dans des situations chimiques qui ne sont pas bien représentées par les bases de données standard. Un premier exemple montre comment la synergie entre l'erreur de délocalisation et les interactions de dispersion reste problématique même pour les fonctionnelles les plus récentes dans les dimères radicalaires chargés asymétriquement. Des solutions sont fournies pour améliorer la description de ces systèmes, qui representent typiquement les porteurs de charge dans les semi-conducteurs organiques. Alors que ce premier chapitre se focalise sur les interactions non-covalentes entre les molécules dans leur état fondamental, on sait très peu de choses sur les conséquences d'un traitement incomplet des interactions de dispersion entre des molécules photo-excitées. En utilisant le stilbène comme exemple de "photoswitch", le deuxième chapitre démontre que négliger ces interactions dans les états excités conduit à des erreurs qualitatives dans la description du processus de photo-déactivation. Les conclusions présentées dans ce chapitre s'appliquent à tout "photoswitch" characterisé par de longues chaînes latérales polarisables.

Contrairement à la chimie quantique traditionnelle (déterministe), les variantes basées sur l'apprentissage automatique sont encore à leurs débuts face au défi de cibler des objets chimiques quantiques fondamentaux, bien que complexes. Les chapitres suivants décrivent le développement et l'application de techniques d'apprentissage automatique pour prédire la densité moléculaire [$\rho(\boldsymbol{r})$] en utilisant une représentation locale et compatible avec la régression du processus gaussien adaptée à la symétrie (SA-GPR). Des applications concrètes sont présentées pour un ensemble chimiquement riche de dimères, dont les densités prédites sont utilisées pour calculer la signature de leurs interactions covalentes et non-covalentes, leurs potentiels électrostatiques, ainsi que leurs énergies d'interaction. La transférabilité du modèle est démontrée par la prédiction de $\rho(\boldsymbol{r})$ pour un ensemble de pentapeptides. Combinant transférabilité et précision, notre modèle de régression permet d'accéder à l'information sur la densité des systèmes chimiques complexes à une fraction du coût des calculs *ab-initio* traditionnels.

Le dernier chapitre exploite la complémentarité des deux stratégies et propose un cadre d'apprentissage automatique capable de quantifier la déviation des fonctionnelles de densité de la condition de linéarité par morceaux de la DFT exacte. Les courbures prédites sont appliquées à la fois pour restaurer la correspondance entre la valeur propre de la HOMO et le premier potentiel d'ionisation et pour fournir une analyse à grande échelle de la relation entre l'écart par rapport à la condition de linéarité par morceaux et les proprités chimiques et structurels.

Dans l'ensemble, les travaux abordés dans cette thèse s'inscrivent dans un effort global visant à étendre l'applicabilité de la KS-DFT à des situations chimiques non-standard et à des systèmes moléculaires de plus en plus complexes tout en tirant parti des dernières avancées de "l'apprentissage machine quantique".

Mots clefs : Chimie Quantique, Théorie de la Fonctionnelle de la Densité, Intelligence Artificielle, Densité Électronique, Erreur de Délocalisation, Interactions Non-Covalentes

Contents

Acknowledgements **i**

Abstract (English/Français) **iii**

List of figures **ix**

1 Introduction **1**

2 Theory **5**
- 2.1 The delocalization error: energy curvature and range-separated functionals . . 5
- 2.2 Atom-pairwise dispersion corrections in Kohn-Sham DFT 8
- 2.3 Numerical Aspects of Kohn-Sham Equations 12
  - 2.3.1 Density Fitting 13
  - 2.3.2 Grid integration of the Exchange-Correlation Potential 15
- 2.4 Gaussian process regression 16
  - 2.4.1 The SOAP representation and similarity kernels 18
  - 2.4.2 Symmetry-adapted Gaussian process regression 20

3 Balancing DFT Interaction Energies in Radical Cation Dimers **23**
- 3.1 Introduction 23
- 3.2 Methods and Computational Details 25
  - 3.2.1 The CryOrel9 set 25
  - 3.2.2 Benchmark level 25
  - 3.2.3 Density functionals and wavefunction based computations 26
  - 3.2.4 Fitting and validation of ωB97X-dDsC 27
- 3.3 Results and discussion 30
  - 3.3.1 CryOrel9 dataset 32
- 3.4 Conclusion 37

4 London dispersion and photochemical processes of molecular switches **39**
- 4.1 Computations Methods 45

5 Transferable Machine-Learning Model of the Electron Density **47**
- 5.1 Introduction 47
- 5.2 Expansion of the electron density into an atom-centered, non-orthogonal basis set 48

5.3 Symmetry-Adapted Gaussian Process Regression for the Electron Density . . . 50
5.4 Results and Discussion . . . 52
5.4.1 Charge decomposition analysis . . . 52
5.4.2 Density learning with SA-GPR . . . 53
5.4.3 Size-extensive extrapolation . . . 55
5.5 Conclusions . . . 56
5.6 Computational Details . . . 57

6 Electron density learning of non-covalent systems 59
6.1 Introduction . . . 59
6.2 Methods . . . 61
6.2.1 Computational Details . . . 62
6.3 Results and Discussion . . . 63
6.3.1 Prediction on Polypeptides . . . 69
6.4 Conclusion . . . 71

7 Learning the energy curvature *versus* particle number 73
7.1 Introduction . . . 73
7.2 Learning Curves . . . 75
7.3 System-specific γ-tuning . . . 78
7.4 Extrapolation . . . 80
7.5 Unsupervised learning and analysis of the QM7 dataset . . . 82
7.6 Conclusion . . . 85
7.7 Computational Details . . . 86

8 General Conclusions and Outlook 87

A CryOrel9 and ωB97X-dDsC 93
A.1 Functional performance on CryOrel. . . . 93
A.2 ωB97X-dDsC . . . 94
A.3 The CryOrel Set . . . 96
A.4 Spin Densities . . . 96

B Energy Curvature: representations and local regression 99
B.1 Performance of different molecular representations in learning the average energy curvature . . . 99
B.2 Local framework for the regression of the energy curvature . . . 100

Bibliography 145

Curriculum Vitae 147

List of Figures

2.1 Lebedev grid with 2030 points on the unit sphere. The color code shows the integration weight of each point. 16

2.2 *a)* Five functions drawn from a GP prior with $\mu = 0$ and a squared exponential covariance function. *b)* Five functions from a GP posterior after data (black points) were given to the model. The black dashed line represents the mean prediction and the gray area is the 2σ confidence interval. 18

3.1 The CryOrel9 set classified by the characteristic crystal packing of each compound a) 2D brickwork, b) columnar-lamellar and c) herringbone packing. The meaning of abbreviations in the Figure is detailed in Appendix A (Table A.2). . . 26

3.2 Main algorithm used for the optimization. 28

3.3 Mean absolute error (MAE) of ωB97X-dDsC on the training set. ωB97X-D is shown for the sake of comparison. 30

3.4 (*left*) MAE of illustrative functionals on the radical cation dimers of Orel26rad and the corresponding neutral compounds of Pi29n.(*right*) MAPE of the same functionals on the Orel26rad dataset. The horizontal line represents the MAPE of ωB97X. 30

3.5 Interaction energy profiles for the radical cation dimers of a) furan and b) thiophene. Insets zoom into the equilibrium region. CCSD(T)/CBS, LC-BOP-LRD and PBE0-dDsC values are taken from the original reference. 32

3.6 Mean absolute error of a) the tested wavefunction based methods and b) the range-separated functionals of the ω-family on the CryOrel9 with respect to DLPNO-CCSD(T)/CBS reference. 33

3.7 Relative and absolute values of U-SAPT0 contributions averaged over CryOrel9 and Orel26rad (left) and divided per type of preferred crystal arrangement in CryOrel9 (right). 34

3.8 (*top*) Correlation between absolute errors of the functionals with respect to DLPNO-CCSD(T)/CBS and the U-SAPT0 (*top left*) exchange and (*top right*) dispersion contributions. Each point represents one dimer of the CryOrel9 dataset. Color code represents different density functionals, while packing motifs are indicated with different symbols. (*bottom*) Spread of the error among the ω-family (standard deviation) for each dimer in CryOrel9. The color code highlights the classification on the basis of the tilt angle, reported on top of each histogram. The inset shows an orthographic view of the ETTDM-TTF dimer. . . 35

3.9 Interaction energy profiles for the radical cation dimers of a) DITT, b) BDT, c) DBT and d) ETTDM-TTF. 36

4.1 Upon optical transition, London dispersion dominates the interaction between: a) unexcited fragments; b) a photoexcited and a ground state molecule; c) two photoexcited molecules. 40

4.2 Structure of the compounds studied: **A** cis-stilbene. **B** cis-3,3'-5,5'-tetra-*tert*-butyl-stilbene. 41

4.3 Energy profiles (in eV) of molecule **A** (left) and **B** (right). Faded lines: S_0 and S_1 energies relative to the S_0 minimum. Solid lines: excitation energies. For the red lines, the D3(BJ) dispersion correction was added to the (TDA-)PBE0/def2-SVP level computations. 43

4.4 Left panel, TDA-PBE0-D3(BJ). Right panel, TDA-PBE0. Comparison between the S_1 energy profiles of molecules **A** (red and orange lines) and **B** (black and blue lines). Purple arrows: barrier to crossing point with the ground state. Green arrows: shift of the S_1 minimum of **B** with respect to **A**. The energies are relative to the respective S_0 minima. 44

4.5 Dynamical evolution of the central dihedral angle (green atoms) and 2-2' carbon-carbon distance (red atoms) on the S_1 surface of the *tert*-butyl substituted stilbene (**B**). Left panel: TDA-PBE0-D3. Right panel: TDA-PBE0. The DHP crossing point region is indicated in blue. The molecular configuration upon the termination of a trajectory is indicated by black dots. 44

5.1 Density errors at different level of representation: (*left*) superposition of isolated atomic densities, (*right*) optimized basis set. Red and blue isosurfaces refer to an error of ±0.005 Bohr^{-3} respectively. The density errors for the structure depicted are reported in the two panels, while the table reports the mean errors over the whole training set for the C_2 and C_2 molecules. 52

5.2 (*top*) representation of the angular momentum decomposition of the electron density. Red and blue isosurfaces refer to ±0.01 Bohr^{-3} respectively. (*bottom*) angular momentum spectrum of the valence electron density of C_2 and C_4 datasets. The isotropic contributions $l = 0$ express the collective variations with respect to the dataset's mean value, while the mean is statistically zero for $l > 0$. 53

5.3 Learning curves for C_2 and C_4 molecules. (*left*) % mean absolute error of the predicted SA-GPR densities as a function of the number of training molecules. The error normalization is provided by the total number of valence electrons. (*right*) root mean square errors of the exchange-correlation energies indirectly predicted from the SA-GPR densities and directly predicted via a scalar SOAP kernel, as a function of the number of training molecules. Dashed lines refer to the error carried by the basis set representation. 54

5.4 Extrapolation results for the valence electron density of one octane (*left*) and one octatetraene (*right*) conformer. (*top*) DFT/PBE density isosurface at 0.25, 0.1, 0.01 Bohr^{-3}, (*middle*) machine-learning prediction isosurface at 0.25, 0.1, 0.01 Bohr^{-3}, (*bottom*) machine-learning error, red and blue isosurfaces refer to ± 0.005 Bohr^{-3} respectively. Relative mean absolute errors averaged over 100 conformers are also reported for both cases. 55

6.1 (*left*) Decomposition error of the electron density of a single water molecule: evolution of the absolute percentage error depending on the choice of decomposition basis set. (*right*) Comparison of the density error made with the standard and the RI-auxiliary cc-pVQZ basis set (cyan and orange isosurfaces refer to an error of ± 0.005 Bohr^{-3}). Reference density: PBE/cc-pVQZ. 63

6.2 Ternary diagram representation of the attractive components of the dimer interaction energies for the 2291 systems considered in this chapter. The values of the SAPT analysis are taken from the original reference. 64

6.3 Learning curves with respect to RI-expanded densities (ML error). (*left*) weighted mean absolute percentage error ($\epsilon_\rho(\%)$) of the predicted SA-GPR densities as a function of the number of training dimers. The weights correspond to the number of electrons in each dimer and the normalization is defined by the total number of electrons. Color code reflects the number of reference environments. (*right*) $\epsilon_\rho(\%)$ of the predicted SA-GPR densities (M=1000) divided per dominant contribution to the interaction energy according to the original reference. . . . 65

6.4 DORI maps of representative dimers for each type of dominant interaction (DORI isovalue: 0.9). Isosurfaces are color-coded with $sgn(\lambda_2)\rho(\boldsymbol{r})$ in the range from attractive -0.02 a.u. (red) to repulsive 0.02 a.u. (blue). In particular, $sgn(\lambda_2)\rho(\boldsymbol{r}) < 0$ characterizes covalent bonds or strongly attractive NCIs (*e.g.* H-bonds); $sgn(\lambda_2)\rho(\boldsymbol{r}) \sim 0$ indicates weak attractive interactions (van der Waals); $sgn(\lambda_2)\rho(\boldsymbol{r}) > 0$ repulsive NCIs (*e.g.* steric clashes). 66

6.5 Electrostatic potential (ESP) maps of representative dimers for each type of dominant interaction (density isovalue: 0.05 e$^-$ Bohr^{-3}). ESP potential is given in Hartree atomic units (a.u.). 67

6.6 (*left*) Electrostatic potential maps 3.25 Å above the plane of the tryptophan (TRP) side-chain. The van der Waals volume of TRP is represented in transparency. The color code represents the electrostatic potential in kcal/mol according the scale chosen in the original reference. (*Right*) Stacking interaction energies of TRP with the phenylalanine (PHE), tyrosin (TYR) and tryptophan (TRP) side-chains computed as detailed in the original reference on the basis of *ab-initio* (*top*) and ML-predicted (*bottom*) ESP. 68

6.7 Weighted mean absolute percentage error (ϵ_ρ(%)) with respect to ωB97X-D/cc-pVQZ densities of the predicted densities extrapolated for 8 biologically relevant peptides (protein databank ID). 69

6.8 *(top left)* predicted electron density of enkephalin (PBD ID: 4OLR) at three isovalues: 0.5, 0.1, and 0.001 e^- Bohr^{-3}. For each isosurface, the L(a,a') similarity index with respect to *ab-initio* density is reported. *(top right)* DORI map of enkephalin (DORI isovalue: 0.9) colored by $sgn(\lambda_2)\rho(\boldsymbol{r})$ in the range from -0.02 a.u. (red) to 0.02 a.u. (blue) *(lower left)* density difference between predicted and *ab-initio* electron density (isovalues ± 0.01e^- Bohr^{-3}). *(lower right)* density difference between predicted and *ab-initio* electron density of 3WNE (isovalues ± 0.01e^- Bohr^{-3}). 70

7.1 Learning curves of the average energy-curvature (C^N_{avg}) as a function of the training set size. The learning exercise is reported for three functionals and Hartree-Fock (HF) using the def2-SVP basis set. The error bars correspond to the standard deviation of the 10-fold cross-validation. The models were built using the SLATM molecular representation. In the upper panels, we report the distribution of C^N_{avg} across the QM7 database at HF and PBE0. 76

7.2 MAE of the model at a training set size of 5000 molecules as a function of the standard deviation of C^N_{avg} using the three functionals and HF. 76

7.3 *(left)* Standard deviations of N-HOMO and (N-1)-LUMO through QM7 with three functionals and HF. *(right)* Pearson's correlation coefficient between the N-HOMO and (N-1)-LUMO energies at different levels of theory. 77

7.4 Schematic representation of the regression framework for the prediction of the optimal range separation parameter per compound. For each molecule, nine independent models predict the energy curvature at LC-ωPBE at nine γ values ranging from 0.1 to 0.9 Bohr^{-1}. The system-specific optimal γ parameter, for which $C^N_{avg} = 0$, is then found by a cubic spline interpolation. 78

7.5 Absolute error between $-\epsilon_{HOMO}$ of LC-ωPBE and its γ-tuned variant and the ionization potential at IP-EOM-CCSD across the 640 molecules of the test set. Optimal γ values derive from the model described in Figure 7.4. The height of the histogram represents the mean absolute error, while the bars show the maximum and the minimum deviations. 79

7.6 Distribution of the optimal γ parameters [Bohr^{-1}] across the 640 molecules of the test set as predicted by the model described in Figure 7.4. The red line show the value of the range-separation parameter in the standard LC-ωPBE. 80

7.7 Extrapolation: optimal γ parameter derived from the model described in Figure 7.4 for two large molecules relevant for the field of hole-transporting materials. The value of $-\epsilon_{HOMO}$ for both the standard LC-ωPBE and its γ-tuned variant are reported along with the reference ionization potentials. 81

7.8 Two-dimensional t-SNE map of the QM7 database on the basis of the SLATM representation. Each point represents a compound colored by its average energy curvature computed with PBE/def2-SVP. The diverging color map highlights the data with the highest and the lowest average energy curvature. Each of the clusters contains molecules with similar patterns, which are defined by the corresponding numbering. 83

7.9 Correlation between the average energy curvature at PBE/def2-SVP and the size of the molecules. The mean values (dots) are computed averaging C^N_{avg} over all the compounds with the same number of non-hydrogen atoms. The error bar represents one standard deviation from the mean. The inset shows the average energy curvature of all the compounds in QM7 with 3 non-hydrogen atoms. The color code in the inset highlights the presence of oxygen (red), nitrogen (blue) or carbon only compounds (black). 84

A.1 MAE of tested functionals on the CryOrel dataset. Blue: GGAs; pale orange: meta-GGAs; pink: global hybrids; cyan: meta-hybrids. Range-separated hybrids follow the same color code as in previous Figures. Oblique bars are drawn to mark the same functional used with a different dispersion correction. 93

A.2 Spin-densities (isovalue: +0.01 e^{-}Bohr^{-3}) of DBT-Sulfone (stacked class), BBBT (tilted class) and ETTDM-TTF (stacked, but asymmetric) with two functionals (ωB97X-dDsC and ωB97M-V), representative of the lower and the upper bound of the interaction energy error. Red asterisks mark which monomer has been optimized as a radical cation. 97

B.1 Learning curves of the average energy-curvature (C^N_{avg}) at LC-$omega$PBE/def2-SVP in function of the training set size. The learning exercise is reported for four different molecular representations: the Coulomb matrix (CM), the Bag of Bonds (BoB), the spectrum of London and Axilrod-Teller-Muto (SLATM) and the smooth overlap of atomic positions (SOAP). The error bars correspond to the standard deviation of the 10-fold cross-validation. 99

B.2 Schematic representation of the local framework for the regression of the energy curvature. 100

1 Introduction

The longstanding goal of quantum chemistry is to establish a coherent relationship between the structure and composition of a molecule and its electronic properties. In principles, there is no constraint in the way this connection can be drawn. From a physical perspective, this relation takes the form of the electronic structure problem, that is the task of determining the state of motion of electrons in the field generated by a set of atomic nuclei.[1,2] Alternatively, given a sufficiently high amount of data, the same connection can be drawn statistically using non-linear regression approaches such as kernel-based machine-learning and artificial neural networks.[3] Despite being equally valid, the deterministic and the statistical approach to quantum chemistry differ dramatically in their stage of development and offer, in consequence, challenges of different nature. The open questions and the interplay between these complementary perspectives on quantum chemistry are the objects of interest of this thesis.

Traditionally, chemical information has been accessed using the deterministic approach, which relies on the development and application of a hierarchy of physically motivated approximations to the exact solution of the electronic Schrödinger equation. Conjugating a low computational cost with an ever-increasing accuracy, Kohn-Sham density functional theory (KS-DFT)[4,5] has become one of the most successful deterministic frameworks for the solution of the electronic structure problem. Through the careful and systematic exploration of the chemical space, approximate KS-DFT has experienced in the last two decades an outstanding evolution,[6–11] which resulted in the identification and development of corrections for its major pitfalls: the delocalization error,[12–17] the incomplete description of London dispersion interactions[18–21] and the single-reference nature of the Kohn-Sham determinant.[4,7,10,17,22,23] Despite this general progress, some chemical situations still represent a challenge for the accuracy of commonly used approximations and correction schemes. These problematic cases are the object of interest in the first part of this thesis. A particular focus is dedicated to the synergistic interplay between delocalization error and London dispersion interactions in asymmetrically charged radical cation dimers and to the analysis of the effects of (missing) London dispersion interactions beyond the standard ground-state situation.

In contrast to Kohn-Sham density functional theory, kernel-based machine-learning and artificial neural networks in quantum chemistry are quite at an early stage of their evolution. Nonetheless, the applications of these frameworks are currently thriving, with the development of predictive models for countless molecular properties. Once trained, these models are orders of magnitude faster to evaluate than traditional *ab-initio* computations, allowing the exploration of otherwise unimaginably vast chemical spaces,[24–31] the access to complex chemical properties at a fraction of the usual computational cost[32–35] and the pursuit of statistically converged results without the need to sacrifice quantum chemical accuracy.[36–40] The unbound potential of artificial intelligence applications in quantum chemistry is currently limited only by the large amount of data needed to construct reliable predictive models. In this sense, kernel-based methods are particularly appreciated in the community as they are the most compatible with a limited number of training instances.[41,42] The evolution of the predictive power of machine-learning in the last few years correlates with the mathematical complexity of the targeted molecular properties, which has grown from simple scalar (*e.g.*, atomization and isomerization energies)[25,41,43] to vectors and tensorial quantities (*e.g.*, forces,[37,44,45] multipole moments,[46] (hyper-)polarizabilities[47,48]) up to complex functions and fields such as potential energy surfaces,[49–52] the electron density[32–34,53–55] and many-body wavefunction.[35] Among all these properties, the electron density is a compelling target for non-linear regression, since it formally contains the same information as the many-body wavefunction, but it is also simply connected with real-space coordinates and properties.[56] Therefore, the second part of this thesis focuses on the construction of a machine-learning model of the electron density. The validity of the density predictions is further demonstrated on prototypical applications such as the topological analysis of bonding and intermolecular interactions and the treatment of electrostatic interactions.

Finally, the third and last objective of this thesis aims at connecting the two approaches and demonstrating how non-linear regression and unsupervised learning become efficient tools to analyze, understand and correct fundamental limitations of approximate functionals. In fact, the deterministic and the statistical approach to quantum chemistry are not mutually exclusive and can largely benefit one from the other. For instance, Müller, Burke and coworkers have demonstrated in a recent landmark work that machine-learning can be used to approximate the kinetic energy functional and its derivatives for one-dimensional, noninteracting fermions.[57] In a complementary example, it is shown that the accuracy of the same machine-learning model can be improved by imposing constraining conditions derived for exact density functional theory.[58] Moreover, the large majority of modern exchange-correlation functionals are developed by combining physically motivated Ansätze with the fitting of adjustable parameters against datasets of accurate molecular properties (for a comprehensive overview see Ref. 11). This deep-rooted practice constitutes further evidence of the historical complementarity of practical KS-DFT and statistical inference.

The material of the thesis is organized as follows.

An overview of the relevant theoretical background is presented in **Chapter 2**. We first introduce the range-separation of the two-electron potential and atom-pairwise dispersion corrections as commonly used strategies to overcome the delocalization error and to account for van der Waals interactions in approximate density functionals. A separate section discusses two practical strategies used to efficiently solve the Kohn-Sham equation that have been crucial to represent the electron density for machine-learning applications. Finally, we summarize the theoretical foundations of Gaussian process regression (GPR) and the smooth overlap of atomic positions (SOAP)[59] for the regression of molecular properties.

Chapter 3 reports an example of a class of molecules that represents a serious challenge even for modern density functionals and correction schemes: the π-dimer radical cations. In the first part of the chapter, we describe the development of a jointly fitted, dispersion corrected, range-separated hybrid density functional (ωB97X-dDsC), specifically built to provide the maximum balance between the treatment of long-range London dispersion and reduction of the delocalization error. The performance of ωB97X-dDsC as well as of other modern functionals of the same family is tested against a database of small π-dimer radical cations, Orel26rad.[60] In the interest of assessing more realistic systems, we additionally report the construction of a dataset of large radical cation dimers (CryOrel9), against which we test the performance of density functionals and state-of-the-art wavefunction based methods.

In **Chapter 4**, we analyze the consequences of van der Waals interactions on the properties of chemical systems beyond their electronic ground state. Here, we compare the excited state properties and molecular dynamics computations of the prototypical *cis*-stilbene molecule with its 3-3',5-5'-tetra-*tert*-butyl derivative. While the explicit treatment of London dispersion interactions results in negligible changes for the *cis*-stilbene, we show that these attractive forces have a substantial impact on the energetics and structural evolution of its substituted derivative. In particular, London dispersion interactions impact the outcome of the simulation qualitatively, increasing the number of trajectories leading to the photocyclization product.

Chapter 5 presents the construction of a local machine-learning framework for the non-linear regression of the valence electron density. The accuracy of the model is demonstrated by predicting the electron density of a conformationally diverse dataset of small hydrocarbons. The scalability and the transferability of the model are then shown with the prediction of the valence electron density of octane and octatetraene while training exclusively on smaller hydrocarbons.

A further development of the machine-learning model presented in the previous chapter is shown in **Chapter 6**. Here, we introduce a different Ansatz for the expansion of the electron density in local, atom-centered contributions, enabling the treatment of core electrons. The regression model is then used to obtain qualitative and quantitative insights using the predicted densities in an ensemble of sidechain–sidechain dimers extracted from the BioFragment database (BFDb).[61] The transferability of the model to more complex chemical systems is demonstrated by predicting and analyzing the electron density of a collection of polypeptides.

In **Chapter 7**, we show how the combination of supervised and unsupervised machine-learning techniques can be used to analyze and correct the spurious energy-curvature *versus* particle number in a selection of common density functionals.[13] In the first part, we build the regression model using the average energy-curvature between the neutral and the first radical cation state of 7165 organic molecules taken from the QM7 database. Then, this information is used for the optimal tuning of the range-separation parameter in LC-ωPBE. In the last section, we apply an unsupervised dimensionality-reduction algorithm to find patterns connecting molecular structure and composition with the degree of convexity of the curvature.

Finally, **Chapter 8** completes this thesis by summarizing the main conclusions and presenting possible future developments.

2 Theory

This chapter provides an overview of the theoretical background relevant to the material presented in this thesis. In particular, the first two sections introduce range-separated hybrid functionals and atom-pairwise dispersion corrections as strategies to overcome two of the most evident failures of approximated Kohn-Sham DFT: the delocalization error and the incomplete description of London dispersion. A separate section is dedicated to density-fitting and grid integration, because of their fundamental role in the work presented in Chapters 5 and 6 as practical tools for the efficient decomposition of the molecular electron density. Finally, the last section summarizes the theoretical foundations of Gaussian process regression (GPR) and introduces the smooth overlap of atomic positions (SOAP) as a powerful tool for the regression of both scalar and tensorial molecular properties.

2.1 The delocalization error: energy curvature and range-separated functionals

The piecewise linear behavior of the total electronic energy as a function of the particle number [$E(N)$] is an exact condition of density functional theory.[13,17,62,63] Following the original demonstration,[13] the ground state of a chemical system with $N+\omega$ particles ($0 \leq \omega \leq 1$) at zero temperature is a statistical mixture of two pure states with N and $N+1$ electrons, respectively denoted $|\Psi_N\rangle$ and $|\Psi_{N+1}\rangle$. In order to conserve the total number of particles, the probabilities associated to these states have to be $1-\omega$ for $|\Psi_N\rangle$ and ω for $|\Psi_{N+1}\rangle$. As a consequence, the expectation value of the Hamiltonian operator ($\hat{H}$) applied on the two-state ensemble is given by

$$\langle \hat{H} \rangle_{N+\omega} = (1-\omega) \cdot \langle \Psi_N | \hat{H} | \Psi_N \rangle + \omega \cdot \langle \Psi_{N+1} | \hat{H} | \Psi_{N+1} \rangle = (1-\omega) E_N + \omega E_{N+1}, \tag{2.1}$$

Eq. 2.1 shows that the total energy between two integer-particle points is indeed a linear function of ω with slope $E_{N+1} - E_N$ and intercept E_N. In general, however, this condition is not realized in common density functional approximations (DFAs), which are instead characterized by a convex (approximately parabolic)[64] $E(N)$ curve.[15,17,65–70] The origin of this incorrect behavior has been attributed to the fact that the LDA and GGA exchange-correlation holes always integrate to -1 electron.[65] While this sum rule is exact for integer states and infinite systems,[71] it is generally too negative for fractional electron numbers.[65,72] The resultant deviation of (semi-)local functionals from piecewise linearity causes the over-stabilization of fractional-particle states and the over-delocalization of electron densities. These spurious effects are known under the collective name of *delocalization error* and affect the evaluation of the dissociation limit of molecules and ions,[13,14,66] the interaction energies of heterodimers (*e.g.,* charge-transfer complexes),[15,73,74] the estimation of the molecular polarizability[75,76] and the fundamental gap.[77–80]

In contrast to LDA and GGA functionals, the Hartree-Fock approximation is characterized by a concave $E(N)$ curvature and overly localized electrons.[67,68] Since the seminal work of Becke,[81,82] the inclusion of Hartree-Fock exchange has been widely recognized as an effective way to reduce the delocalization error in (semi-)local functionals.[17,83–86] Although the construction of global hybrids was originally proposed on the basis of the adiabatic connection approach,[87–90] the mutual cancellation of the $E(N)$ curvature offers an alternative perspective on the benefits of mixing.[62,67] Nevertheless, adding a constant fraction of Hartree-Fock exchange through all space violates, yet another exact condition of DFT: the exchange-correlation potential should decay as $-1/r$ for $r \to \infty$, where r represents the distance of an electron from the atomic nucleus.[91–93] As demonstrated by Almbladh and von Barth,[92] an electron escaping from a finite system should experience at large distances the mean-field of the ion it leaves, which corresponds to the Coulomb potential of a single positively charged particle. Although the asymptotic behavior of density functionals is generally incorrect (*e.g.*, LDA has an exponential decay),[93] even the simplest approximations are highly accurate at short inter-electronic distances.[89,94,95] In fact, the short-range exchange-correlation energy can be represented as a local expansion of the on-top pair density [$\Pi(\boldsymbol{r}, \boldsymbol{r})$] and its derivatives, which are respectively exact and nearly exact already in LDA.[96] The accuracy of simple density functionals in inter-electronic regions, where the wavefunction expansion converges slowly,[97,98] was the original motivation proposed by Savin for splitting the two-electron Coulomb operator in a short- and a long-range part.[99,100] The most compelling advantage of range-separation is that it restores the exact asymptotic behavior of density functionals, by gradually increasing the fraction of exact-exchange as the inter-electronic distance increases. Although other propositions exist,[96,99,101] the splitting of the Coulomb potential is typically performed on the basis of a weighted error function [$\mathrm{erf}(\omega \cdot |\boldsymbol{r}_1 - \boldsymbol{r}_2|)$], so that

$$\frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} = \frac{1 - \mathrm{erf}(\omega \cdot |\boldsymbol{r}_1 - \boldsymbol{r}_2|)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} + \frac{\mathrm{erf}(\omega \cdot |\boldsymbol{r}_1 - \boldsymbol{r}_2|)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \tag{2.2}$$

where ω is an adjustable parameter that determines how fast the change from short- to long-range regime occurs, while $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are two electronic coordinates. In principle, the optimal value of this parameter is system dependent[62,99,102,102,103] and should be computed by enforcing the Koopmans'/Janak's theorems or by minimizing the energy-curvature with respect to the number of particles.[104,105] On the other hand, this procedure breaks size-consistency,[106] and therefore the common practice is to keep the ω parameter fixed for a given functional.

The problem of combining system-specific range-separation and size-consistency have been addressed with the introduction of two distinct approximations: local range-separation and range-separated local hybrids.[107] Both methods are extensions of the local hybrid approach (LH),[108] where a local mixing function (LFM) determines the amount of exact-exchange to mix at each point in space (for a recent review on LH, see Ref. 109). In the local range-separation approach, the ω parameter is position-dependent $[\omega(\boldsymbol{r})]$ and it is determined locally on the basis of a gradient expansion of the characteristic correlation length, given by the Wigner-Seitz radius.[110] In contrast, range-separated local hybrids are characterized by a universal ω parameter and a position-dependent admixture of HF-exchange at short-range.[111,112] Similar to local hybrids, these methods are characterized by nonstandard two-electron integrals,[108] whose evaluation hinders their computational efficiency. As a consequence, their use is not widespread and standard range-separated hybrids still represent the most commonly used approach to reduce the effects of the delocalization error in routine computations.

The previous paragraphs presented the problem of the energy curvature in the general case of many-electron systems. In the one-electron limit the deviation from piecewise linearity is also well-defined and it is the manifestation of another limitation of approximate KS-DFT, the *self-interaction error* (1-e SIE).[12,113–118] This term indicates a condition where the two-electron energy of a one-electron system is not zero, as the particle experiences a repulsive Coulomb interaction with itself. While the presence of the self-interaction error was already observed in the earliest times of DFT,[119–121] it was only in 1981 that Perdew and Zunger reported the basic requirement for an arbitrary functional to be one-electron SIE free:[12]

$$E_{ee}[\rho_i^\alpha,0] = \frac{1}{2}\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho_i^\alpha(\boldsymbol{r}_1)\rho_i^\alpha(\boldsymbol{r}_2)}{|\boldsymbol{r}_1-\boldsymbol{r}_2|} - \frac{1}{2}\int d\boldsymbol{r}_1 d\boldsymbol{r}_2\, \rho_i^\alpha(\boldsymbol{r}_1)\frac{\delta E_{XC}[\rho_i^\alpha,0]}{\delta\rho_i^\alpha(\boldsymbol{r}_1)}\rho_i^\alpha(\boldsymbol{r}_2) = \tag{2.3}$$

$$= J[\rho_i^\alpha] + E_{xc}[\rho_i^\alpha,0] = 0 \tag{2.4}$$

where E_{ee} denotes the two-electrons potential energy, $\rho_i^\alpha(\boldsymbol{r}) = |\phi_i(\boldsymbol{r},\alpha)|^2$ are the single-particle densities associated with the i^{th} α-spin orbital $[\phi_i(\boldsymbol{r},\alpha)]$, J is the classical Coulomb repulsion energy as defined by the first integral in Equation 2.3 and the functional derivative $\frac{\delta E_{XC}[\rho_i^\alpha,0]}{\delta\rho_i^\alpha(\boldsymbol{r}_1)}$ represents a general exchange-correlation potential with energy E_{xc}. The same expression applies to β-spin electrons.

By enforcing the equality in Equation 2.4, the exchange-correlation energy of a given functional ($E_{xc}[\rho^{\alpha},\rho^{\beta}]$) is corrected for the one-electron self-interaction as

$$E_{xc}^{SIC} = E_{xc}[\rho^{\alpha},\rho^{\beta}] - \left(\sum_{i,\sigma} J[\rho_i^{\sigma}] + E_{xc}[\rho_i^{\sigma},0]\right). \tag{2.5}$$

Equation 2.5 (PZ-SIC) is exact by construction for one-electron systems and the correction term reduces to zero for the exact exchange-correlation functional. However, since the PZ-SIC depends on single orbital densities (as opposed to the total density), it is not invariant under orbital unitary transformations.[12,122,123] As first suggested in the original work,[12] performing orbital localization prior to correction bypasses the problem, but it simultaneously introduces an undesirable dependence on the chosen localization scheme.[124–126] Another possibility to solve the problem is to use the PZ-SIC in combination with an optimized effective potential method,[127] which, however, increases dramatically the overall computational cost. More recent alternatives to restore the unitary invariance in PZ-SIC include its reformulation on the basis of Fermi orbitals (FO-SIC)[128] or its self-consistent implementation in combination with optimal complex orbitals.[129,130]

In contrast to approximate KS-DFT, the Hartree-Fock method does not suffer from the self-interaction error, as the self-exchange contribution (E_X^{HF}) exactly cancels the self-Coulomb interaction (J) for one electron in orbital $\phi(\boldsymbol{r})$,

$$E_X^{HF}[\phi(\boldsymbol{r})] = -\frac{1}{2}\int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\phi^*(\boldsymbol{r}_1)\phi(\boldsymbol{r}_1)\phi^*(\boldsymbol{r}_2)\phi(\boldsymbol{r}_2)}{|\boldsymbol{r}_1-\boldsymbol{r}_2|} = -J \tag{2.6}$$

Therefore, the electronic energy of the Hartree-Fock method is piecewise linear for one-electron systems, but not in the many-electron case.[67,68] The same argument holds also for density functionals relying on 100% exact exchange (*e.g.*, MCY2[131] and M06-HF[132]), which do not generally perform better than other hybrids for many-electron systems.[15,16,66,118] The $E(N)$ curvature, therefore, represents a more effective metric than the one-electron SIE to assess the overall quality of density functional approximations.

2.2 Atom-pairwise dispersion corrections in Kohn-Sham DFT

As early as the mid-90s, it was recognized that approximate exchange-correlation density functionals yield an inconsistent description of London dispersion interactions.[18–21] Since then, several correction schemes to KS-DFT have been proposed and can be classified into three distinct categories:[133,134] explicit non-local correlation functionals (*e.g.*, vdW-DF[135,136], VV10[137]), effective 1-electron potentials (*e.g.*, DCACP,[138,139] Minnesota functionals[140]), and *ad-hoc* atom-pairwise dispersion corrections. The last family of corrections qualifies for a more detailed discussion in this thesis, due to its central role in the Chapters 3 and 4.

Ad-hoc (or semiclassical) [134] atom-pairwise schemes are the most commonly used corrections in routine computations due to their computational efficiency and accuracy. [133,134,141,142] The common feature of these corrections is that the London dispersion energy is computed between atom pairs and added *a posteriori* to the converged KS-DFT computation. Their shared theoretical foundation lies into the evaluation of long-range inter-electronic correlations with a perturbative approach. [134] Within this framework, the order in the perturbation expansion determines the maximum number of interactions considered, *i.e.* pairwise schemes are truncated to second-order (PT2, Eq. 2.7). [143] By including higher terms, the majority of *ad-hoc* dispersion corrections have been extended to account for many-body effects. [144–146]

The general form of pairwise dispersion corrections is obtained by considering the PT2 correlation energy between two molecular fragments (**A**,**B**) placed at infinite distance from one another ($R_{AB} \rightarrow \infty$). In the limit where the electronic excitations of each fragment are localized, the second-order correlation energy (E^{PT2}_{corr}) becomes the square of the Coulomb interaction between two transition densities: [134]

$$\lim_{R_{AB}\rightarrow\infty} E^{PT2}_{corr} = E^{AB}_{disp} = -\sum_{i,a\in A}\sum_{j,b\in B}\frac{(ia|jb)^2}{\omega_{ai}+\omega_{bj}}, \tag{2.7}$$

where ij and ab are the usual indices for occupied and unoccupied orbitals, $(ia|jb)$ is the Coulomb integral written in Mulliken notation, and the two ω are the excitation energies including the effects of the inter-fragment interactions (coupled excitations). Although not strictly needed, [147] the evaluation of Eq. 2.7 is greatly simplified by the expansion of the Coulomb potential in multipole moments, among which the first non-vanishing contribution is given by the dipole-dipole term (μ_{ia}). In the case of spherical atoms, the leading multipole term of Eq. 2.7 is

$$E^{AB}_{disp} = -\frac{3}{2}\sum_{i,a\in A}\sum_{j,b\in B}\frac{|\mu_{ia}|^2|\mu_{jb}|^2}{(\omega_{ai}+\omega_{bj})R^6_{AB}} = -\frac{C^{AB}_6}{R^6_{AB}}, \tag{2.8}$$

where C^{AB}_6 is the dipole-dipole dispersion coefficient. Inclusion of higher multipole interactions (dipole-quadrupole, dipole-octopole, quadrupole-quadrupole, *etc.*) add to Eq. 2.8 additional terms with a faster radial decay (R^{-8},R^{-10}, *etc.*).

For practical applications, Equation 2.8 and its extensions have two main drawbacks. First, the dispersion coefficients (C^{AB}_n, with n=6,8,10...) can be derived from experimental dipole oscillator strengths [148] or computed using frequency-dependent polarizabilities, [144,149,150] but they are rarely known for any two arbitrary molecular fragments. As common practice in force field applications,[151] the general strategy to overcome this limitation is to compute the inter-fragment C_n^{AB} as a pairwise sum of their atomic dispersion coefficients, scaled to account for their dependence on the molecular environment. Second, the multipole expansion of the Coulomb potential is only well-defined at large inter-fragment distances and diverges for $R_{AB} \rightarrow 0$. The singularities of the Eq. 2.8 are avoided in practice using a damping function (f_{damp}^n), which goes to zero[148,152–156] or to a finite negative value[147,157–159] for $R_{AB} \rightarrow 0$. Including these two practical expedients, the most general expression of the London dispersion energy in semiclassical atom-pairwise schemes becomes

$$E_{disp}^{AB} = -\sum_{x \in A} \sum_{y \in B} \sum_{n=6,8,10...} \frac{C_n^{xy}}{R^n} f_{damp}^n, \tag{2.9}$$

where x and y are atomic indices.

Modern atom-pairwise corrections can be regrouped into five different families: Grimme's DFT-D method (D2,[154] D3,[144] D3(BJ)[160] and recently D4[161]), the Tkatchenko-Scheffler vdW-TS scheme,[162] the approaches based on maximally localised Wannier functions (vdw-WF),[163–165] those based on Becke-Johnson's exchange-hole dipole moment (XDM[157–159,166–168] and dDsC[169–171]) and the local response dispersion method (LRD)[172,173]. Each of these corrections is different from the others according to the particular choice of the damping function, order of truncation (n) and the way the dispersion coefficients and their dependence on molecular environments are evaluated.

Among all the mentioned schemes, we present a more detailed description of the dDsC dispersion correction, due to its central role in Chapter 3. The dispersion coefficients in dDsC are computed within the framework of the Becke-Johnson exchange-dipole moment (XDM).[157–159,166–168] The underlying idea of the formalism is that an electron and its exchange hole are characterized by nonvanishing multipole moments, whose mutual interactions with the exchange and induced multipole moments of another fragment are responsible for London dispersion forces.[166,168] The description of the exchange hole is, therefore, a crucial aspect of the XDM formalism. In the original work,[166] Becke and Johnson used the exact nonlocal expression for the hole and later introduced a linear-scaling variant[157] based on the local Becke-Roussel (BR) formalism.[174] By construction, the BR model reproduces the properties of the exact exchange hole up to second-order in a local Taylor expansion around a reference point[174] and depends, as a consequence, on the electron density [$\rho(\boldsymbol{r})$], as well as its curvature [$\nabla^2\rho(\boldsymbol{r})$] and the local kinetic energy density [$\sum(\nabla\psi(\boldsymbol{r}))^2$]. In contrast, the dDsC dispersion correction uses a reformulation of the exchange hole with reduced numerical complexity, as it is only based on the electron density and the reduced density gradient.[170] In dDsC, the square of the exchange-hole dipole moment for spin σ [$\mu^2_{X,\sigma}(\boldsymbol{r})$] takes the following form:

$$\mu^2_{X,\sigma}(\boldsymbol{r}) = \left(A \cdot s \cdot r_s e^{-B \cdot s}\right)^2, \tag{2.10}$$

where $s = \frac{|\nabla\rho(\boldsymbol{r})|}{2(3\pi^2)^{1/3}\rho^{4/3}(\boldsymbol{r})}$ is the reduced density gradient, $r_s = \left(\frac{3}{4\pi\rho(\boldsymbol{r})}\right)^{1/3}$ is the local Wigner-Seitz radius, A=2 and B=1 are two fixed parameters fit on rare gas homodimers.[170] As in the original XDM method, the atomic partitioning of $\mu^2_X(\boldsymbol{r})$ is performed as:

$$\langle\mu^2_{X,A}\rangle = \sum_\sigma \int d\boldsymbol{r}\, w_A(\boldsymbol{r})\rho_\sigma(\boldsymbol{r})\mu^2_{X,\sigma}(\boldsymbol{r}). \tag{2.11}$$

In contrast to XDM (and vdW-TS), however, the atomic weights $w_A(\boldsymbol{r})$ are computed in dDsC with the classical Hirshfeld-dominant partitioning, which is a binary scheme that attributes full weight ($w_A(\boldsymbol{r}) = 1$) only to the atom with the largest classical Hirshfeld contribution. This variant of the Hirshfeld scheme retains the numerical ease of the classical partitioning and simultaneously defines well-localized atomic basins, which are naturally compatible with exchange hole multipole expansion.[175]

Using Eq. 2.11, the dispersion coefficients for two atoms are computed via a modified Slater-Kirkwood formula[166,176,177] as

$$C_6^{AB} = \frac{\alpha_A^0\alpha_B^0 \langle\mu^2_{X,A}\rangle\langle\mu^2_{X,B}\rangle}{\alpha_B^0\langle\mu^2_{X,A}\rangle + \alpha_A^0\langle\mu^2_{X,B}\rangle} \tag{2.12}$$

where α_A^0 and α_B^0 are the atom-in-molecule static polarizabilities, estimated scaling the free atomic polarizability with the ratio between the atom-in-molecule and the free atomic volume.[178] Higher-order (C_8 and C_{10}) dispersion coefficients can be evaluated by generalizing the above description to the exchange-quadrupole and octupole moments.

The damping of the dDsC correction is based on the universal Tang and Toennies damping function[152,153]

$$f_{2n}(x) = 1 - \exp(-x)\sum_{k=0}^{2n}\frac{x^k}{k!}. \tag{2.13}$$

The damping argument x in Eq. 2.13 is defined as $x = bR_{AB}$, where R_{AB} is the distance between two atoms and b is an additional damping factor. In contrast to the original formulation where b is only a fitted parameter,[152] the damping factor in dDsC is itself a function of the interatomic distance and contains a second Fermi-like function $\left(F(x) = \frac{2}{e^{a_0 \cdot x}+1}\right)$ that controls the behavior of the correction at short distances (Eq. 2.14):[171]

$$b_{dDsC}(R_{AB}) = F(x) \cdot b(R_{AB} \to \infty). \tag{2.14}$$

The argument of the Fermi-like function is defined as

$$x = \left(2q_{AB} + \frac{|(Z_A - N_A^D)(Z_B - N_B^D)|}{R_{AB}}\right) \frac{N_A^D + N_B^D}{N_A^D \cdot N_B^D}, \tag{2.15}$$

where Z_A is the nuclear charge of atom A, N_A^D its Hirshfeld dominant population, and $q_{AB} = \int d\boldsymbol{r}\, w_A(\boldsymbol{r}) w_B(\boldsymbol{r}) \rho(\boldsymbol{r})$ is a covalent bond index based on the overlap of Hirshfeld classical populations of atoms A and B $[w_A(\boldsymbol{r})$ and $w_B(\boldsymbol{r})]$. Using the preceding definitions, it follows that $F(x) = 1$ for $R_{AB} \to \infty$ and $F(x) = 0$ for $R_{AB} \to 0$. The second term, $b(R_{AB} \to \infty)$, denotes the asymptotic value of the damping parameter that is estimated on the basis of the effective atom-in-molecule polarizabilities.[171] The dDsC damping function is controlled by two functional dependent parameters, one scaling the steepness of the Fermi function and the other adjusting the value of $b(R_{AB} \to \infty)$. Combining an efficient evaluation of dispersion coefficients with a double damping function, dDsC offers enough flexibility to accurately describe electron correlation effects at all ranges, from the intra- to the inter-molecular domain.

2.3 Numerical Aspects of Kohn-Sham Equations

The exact ground-state electronic energy of a chemical system (E_0) can be expressed on the basis of the Hohenberg-Kohn theorems[56] and their extension to N-representable[a] densities[179] as:

$$E_0 = \min_{\rho \to N} \left(\min_{\Psi \to \rho} \langle \Psi | \hat{T} + \hat{V}_{ext} + \hat{V}_{ee} | \Psi \rangle \right), \tag{2.16}$$

where the constrained search is first performed over all the antisymmetric wavefunctions (Ψ) that yield a fixed trial density (ρ) and then over all the trial densities that integrate to the correct number of electrons (N). $\hat{T}$, $\hat{V}_{ext}$ and $\hat{V}_{ee}$ are, respectively, the kinetic energy, the

---

[a] An N-representable density is any density that can be obtained from an antisymmetric wavefunction.

external potential (electron-nuclei) and the electron-electron (see Eq. 2.3) operators. In 1965, Kohn and Sham proposed a practical reformulation of the original DFT framework, where the fully interacting reference system was replaced by an artificial set of non-interacting electrons with the same ground-state density.[4] The exact wavefunction of the non-interacting system simply reduces to a single determinant (Ψ_s), whose kinetic energy (T_s) is readily evaluated as:

$$T_s[\rho] = \min_{\Psi_s \to \rho} \langle \Psi_s | \sum_i^N -\frac{1}{2}\nabla_i^2 | \Psi_s \rangle = \langle \Psi_s[\rho] | \sum_i^N -\frac{1}{2}\nabla_i^2 | \Psi_s[\rho] \rangle \,. \tag{2.17}$$

Using Equation 2.17, the total electronic energy in Kohn-Sham DFT can be written as

$$E[\rho] = T_s[\rho] + E_{\text{ext}}[\rho] + J[\rho] + [(T[\rho] - T_s[\rho]) + (V_{ee}[\rho] - J[\rho])] = \tag{2.18}$$

$$= T_s[\rho] - \sum_A Z_A \int d\boldsymbol{r} \frac{\rho(\boldsymbol{r})}{|\boldsymbol{r} - \boldsymbol{R}_A|} + \frac{1}{2} \int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} + E_{XC}[\rho]. \tag{2.19}$$

The above expression defines the external potential energy ($E_{\text{ext}}[\rho]$, Z_A and R_A being the charge and position of nucleus A), the classical electron-electron repulsion ($J[\rho]$), as well as the exchange-correlation energy functional ($E_{XC}[\rho]$), whose exact form and analytic dependence on the density remain yet unknown.

Any routine Kohn-Sham DFT computation relies on practical numerical strategies to evaluate the different terms in Equation 2.19. These backbone tools are both physically motivated approximations and numerical schemes that dictate the overall scaling and tractability of the computations. Among these techniques, which range from SCF accelerators[180–189] to the generation of an initial guess for the density[190–194] to the numerical evaluation of integrals,[195–199] the density-fitting approximation and the grid integration of the exchange-correlation functional are particularly relevant for the work presented in this thesis.

2.3.1 Density Fitting

The development of fitting techniques to reduce the computationally demanding evaluation of four-center two-electron integrals (*e.g.*, Eq. 2.20) has been an active field of research since the early days of quantum chemistry.[200,201]

$$(ab|cd) = \int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \, \frac{\chi_a(\boldsymbol{r}_1)\chi_b(\boldsymbol{r}_1)\, \chi_c(\boldsymbol{r}_2)\chi_d(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} = \int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho_{ab}(\boldsymbol{r}_1)\rho_{cd}(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \tag{2.20}$$

Density fitting (DF) is an alternative approach to the exact evaluation of Eq. 2.20, which is widely used to approximate the Coulomb and exchange terms in HF and DFT,[202–204] as well as to accelerate integral transformations in *post*-HF methods.[205–211] The technique consists of approximating the binary products of atomic orbitals $[\chi_a(\boldsymbol{r})\chi_b(\boldsymbol{r})]$ by an expansion on an auxiliary basis set $\phi^{\mathrm{Aux}}(\boldsymbol{r})$,[207,212–215] such as

$$\rho_{ab}(\boldsymbol{r}) = \chi_a(\boldsymbol{r})\chi_b(\boldsymbol{r}) = \sum_{P}^{N_{aux}} d_P^{ab}\phi_P^{\mathrm{Aux}}(\boldsymbol{r}) = \bar{\rho}_P(\boldsymbol{r}). \tag{2.21}$$

The expansion coefficients (d_P^{ab}) are obtained by minimizing a functional of the error between the approximated $[\bar{\rho}_P(\boldsymbol{r})]$ and the *ab-initio* $[\rho_{ab}(\boldsymbol{r})]$ one-electron densities. Originally derived by Whitten in the context of *ab-initio* correlated methods[214] and subsequently popularized for DFT applications by Dunlap, Connolly and Sabin,[215] the most widely used functional has the form of a Coulomb integral:

$$\varepsilon_{ab} = \int d\boldsymbol{r}_1 \int d\boldsymbol{r}_2 \frac{[\rho_{ab}(\boldsymbol{r}_1) - \bar{\rho}_P(\boldsymbol{r}_1)][\rho_{ab}(\boldsymbol{r}_2) - \bar{\rho}_P(\boldsymbol{r}_2)]}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|}. \tag{2.22}$$

In other (less common) formulations of Eq. 2.22, the Coulomb potential is replaced by a different operator, including overlap $(\delta(\boldsymbol{r}_1 - \boldsymbol{r}_2))$,[202,213,216] anti-Coulomb $(-|\boldsymbol{r}_1 - \boldsymbol{r}_2|)$[217] and, more recently, the attenuated-Coulomb operator $\left(\frac{1-\mathrm{erf}(\omega|\boldsymbol{r}_1-\boldsymbol{r}_2|)}{|\boldsymbol{r}_1-\boldsymbol{r}_2|}\right)$.[218]

Minimizing the Coulomb energy of the fitting residuals according to Eq. 2.22 leads to the following definition of the coefficients:

$$d_P^{ab} = \sum_Q (ab|Q)[\boldsymbol{J}^{-1}]_{QP} \tag{2.23}$$

where $(ab|Q)$ and $[\boldsymbol{J}]_{QP}$ are respectively three- and two-center electron repulsion integrals. Using Eqs. 2.21 and 2.23, the usual four-center two-electrons integrals (Eq. 2.20) takes the following form:

$$(ab|cd) = \sum_{PQ}^{N_{aux}} (ab|Q)[\boldsymbol{J}^{-1}]_{QP}(P|cd) \tag{2.24}$$

where the summation is performed over the auxiliary basis set. As the final expression (Eq. 2.24) only depends on two- and three-centers integrals, the overall scaling of its evaluation is reduced to $O(N^3)$ from the formal $O(N^4)$, where N is the number of atomic orbitals. This result makes density fitting particularly useful for computations involving large molecules and basis sets. Moreover, avoiding four-center integrals accelerates dramatically the integral transformations from the atomic to the molecular orbitals basis,[211] which is a fundamental advantage to contain the overall cost of *post*-HF methods. Recent developments in the density fitting approach, such as the MARI-J method of Ahlrichs and coworkers[219] or the CFMM-DF technique developed in the Head-Gordon group,[220] aim at further reducing the $O(N^3)$ scaling and reaching the final goal of a linear scaling evaluation of electron repulsion integrals.

2.3.2 Grid integration of the Exchange-Correlation Potential

All the integrals needed to compute the Kohn-Sham electronic energy (see Eq. 2.19) can be evaluated analytically, with the only exception of those involving the exchange-correlation potential $[V_{XC}(\boldsymbol{r})]$.[221] The complexity of this type of integrals can be effectively shown using the simplest approximation for an exchange functional (Dirac's exchange):[222]

$$E_X[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int_{-\infty}^{\infty} d\boldsymbol{r}\ \rho(\boldsymbol{r})^{4/3}. \tag{2.25}$$

Only particularly simple forms of $\rho(\boldsymbol{r})$ (*e.g.*, uniform electron gas or a single Gaussian function in the minimal basis hydrogen atom) allows the analytical evaluation of the above expression. Therefore, numerical integration is unavoidable in the majority of cases and KS-DFT computations commonly require an integration grid. The characteristics of the real-space grid, such as its distribution and size, are essential to ensure a good trade-off between computational efficiency and high numerical accuracy.[223] Numerical quadrature, in fact, does not alter the overall scaling of KS-DFT computations, but it does add a large prefactor dependent on the total grid size.[224] As the exchange-correlation potential is mainly governed by the behavior of the electron density (*e.g.*, Eq. 2.25) and its derivatives, the most effective quadrature grids are constructed according to the features of the density, such as its accumulation and loss of its angular structure approaching the nuclei.[195]

While several quadrature schemes exist for molecules, their large majority is based on the decomposition of the molecular integrals into a sum of weighted atomic contributions, where the atomic domains are usually obtained by dividing the molecule in Voronoi polyhedra[225] or in overlapping continuous cells.[195] The integration over each atomic domain is further divided in radial and angular parts. Radial integration is commonly performed by transforming the integration limits ($0 \leq r < \infty$) to a finite interval, often $0 \leq x < 1$ or $-1 \leq x < 1$. Then, the quadrature is performed using a Gauss(-Chebyshev) scheme[195,223] or using the Euler-MacLaurin formula for a chosen set of radial points.[196,226]

In contrast to radial integration, where many equally efficient schemes are available,[227] angular quadrature is mostly performed using Lebedev grids (Figure 2.1).[228–232] This scheme

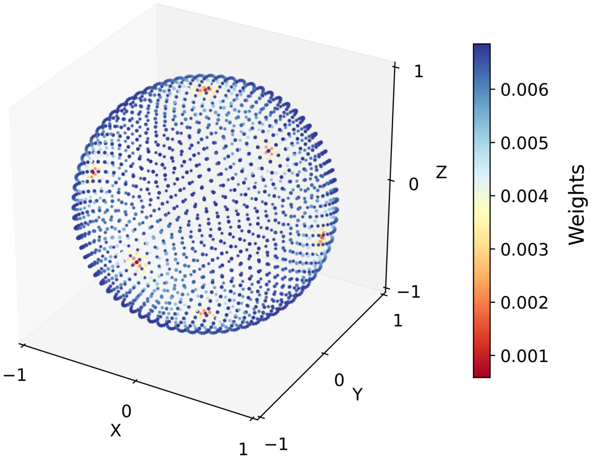


Figure 2.1 – Lebedev grid with 2030 points on the unit sphere. The color code shows the integration weight of each point.

samples the unit sphere with grid points taken from an inscribed octahedron, starting from the 6 vertices, continuing with the 8 and 12 centers of the faces and sides, adding points until the desired density is reached. The weight of each point is defined in such a way that the grid exactly integrates all spherical harmonics up to order $L \sim \sqrt{3n} - 1$, where n is the number of grid points.[224] In addition, Lebedev grids are invariant under the rotations of the octahedral group and under inversion, which facilitates the use of molecular symmetries to reduce the number of angular points, as practically all the common molecular point groups are directly subgroups of O_h or have large subgroups in common with it.

Besides Lebedev grids, angular quadrature can be performed directly using products grids of the spherical polar coordinates, using a Gauss-Legendre scheme for θ and a simple Gauss scheme for ϕ.[196] Nevertheless, this methodology is much less efficient as it requires $\frac{3}{2}$ more points than the Lebedev integration to integrate all the spherical harmonics up to the same order.[224]

2.4 Gaussian process regression

The original paradigm of quantum chemistry relies upon the construction of a hierarchy of ever more accurate physical approximations to reach the exact solution of the Schrödinger equation. Given these models and a molecule, it is possible to compute any chemical property, with an accuracy dependent on the degree of approximation of the chosen model. Mathematically, this procedure can be expressed as a given set of molecular variables (*e.g.*, atomic positions, charges, spin-states, *etc.*) $\{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}\}$ to be mapped onto a set of molecular properties $\{\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_N}\}$, through a known function $f : \boldsymbol{X} \rightarrow \boldsymbol{Y}$.

The core concept behind machine-learning techniques, also valid for quantum chemical applications, is to invert this relation and given a set of variables and their respective properties $\{(\boldsymbol{x_1}, \boldsymbol{y_1}), (\boldsymbol{x_2}, \boldsymbol{y_2}), ..., (\boldsymbol{x_N}, \boldsymbol{y_N})\}$ infer the mapping f, in order to be able to predict the property ($\boldsymbol{y_a}$) for any new input value ($\boldsymbol{x_a}$).[233–235] In contrast to traditional quantum chemistry, where the form of the mapping is uniquely determined by the physical approximations made, any function compatible with the data would be equally valid in the inductive framework of machine-learning.[236] Since only a finite number of points are known, there is always an infinite set of functions that satisfies all the relations $\boldsymbol{X} \rightarrow \boldsymbol{Y}$. This is an impasse for which a step forward is to assume some of the characteristics (*e.g.,* the smoothness) of the mapping f and to attribute to each function a prior probability on this basis. Although in this way some functions are much more likely than others, they are still infinitely many.

The key to solve this issue is to consider the infinite set as a stochastic distribution of continuous functions (*i.e.,* the distribution of a stochastic *process*), rather than as individual units. Within a Bayesian framework,[237,238] the prediction of the properties ($\boldsymbol{y_a}$) of new inputs ($\boldsymbol{x_a}$) can be derived from the probability distribution over the functions given the known data [$p(f|\boldsymbol{X}, \boldsymbol{Y})$] as:

$$p(\boldsymbol{y_a}|\boldsymbol{x_a}, \boldsymbol{X}, \boldsymbol{Y}) = \int df\, p(f|\boldsymbol{X}, \boldsymbol{Y})\, p(\boldsymbol{y_a}|\boldsymbol{x_a}, f). \tag{2.26}$$

The evaluation of the terms in Equation 2.26 is greatly simplified if the stochastic process is a Gaussian process (GP), *i.e.*, if any finite collection of its random variables is normally distributed.[239] In fact, akin to a standard Gaussian distribution, only the first and the second moment (*i.e.,* the mean [$\mu(\boldsymbol{X})$] and covariance [$\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X'})$] functions) are necessary to completely determine a Gaussian process.[240] It follows that if the Bayesian prior for f is let to be a GP, then the posterior $p(\boldsymbol{y_a}|\boldsymbol{x_a}, \boldsymbol{X}, \boldsymbol{Y})$ can be written in terms of a multivariate Gaussian distribution:[236]

$$p(\boldsymbol{y_a}|\boldsymbol{x_a}, \boldsymbol{X}, \boldsymbol{Y}) = \mathscr{N}(\boldsymbol{y_a}|\mu^*, \boldsymbol{K}^*) \tag{2.27}$$

where,

$$\begin{aligned} &\mu^* = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x_a})^T \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{Y} \ \text{ and} \\ &\boldsymbol{K}^* = \boldsymbol{K}(\boldsymbol{x_a}, \boldsymbol{x_a}) - \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x_a})^T \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x_a}). \end{aligned}$$

The choice of the covariance function (or kernel) [$\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X'})$] is a key element of Gaussian process regression (GPR), as it can be selected *ad hoc* on the basis of educated assumptions about the form of the targeted function.[236,241,242] As $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X'})$ measures the covariance of two input values, it can be also considered, with a change of perspective, as a measure of their similarity (or proximity). In general, not any arbitrary function can be a kernel, the condition being that it must be positive semidefinite (i.e., all the eigenvalues of its matrix form are non-negative).[243]

Figure 2.2 shows an example of the use of a Gaussian process for the Bayesian regression of a simple one dimensional function using a squared exponential kernel ($K(X, X') = e^{\frac{|X-X'|^2}{2\sigma^2}}$).[236] Note that the functions in the Figure are only for illustrative purposes, as the mean prediction and the confidence interval are computed with a virtually infinite set of functions.

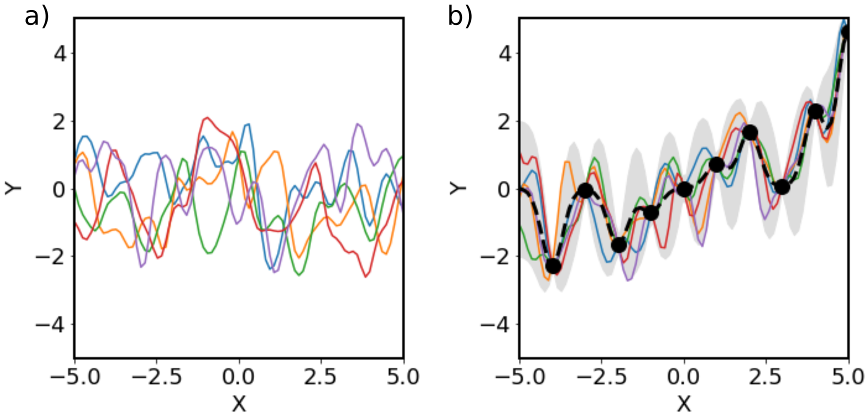


Figure 2.2 – *a)* Five functions drawn from a GP prior with $\mu = 0$ and a squared exponential covariance function. *b)* Five functions from a GP posterior after data (black points) were given to the model. The black dashed line represents the mean prediction and the gray area is the 2σ confidence interval.

2.4.1 The SOAP representation and similarity kernels

Applying the GPR framework to solve quantum chemistry problems implies devising a strategy to encode molecular information into a set of input vectors ($\{\boldsymbol{X}\}$).[25,27,38,59,244–251] For the sake of the regression, these vectors are required to completely determine the targeted property and to encode all its fundamental symmetries.[59,250] For example, the Hamiltonian and, thus, the energy of any given molecule is defined by the number of electrons (N) and the external potential (V_{ext}), which are determined in turn by the position of the nuclei ($\boldsymbol{R_I}$) and their charges (Z_I).[252] Therefore, $\boldsymbol{R_I}$ and Z_I carry sufficient information to build a molecular representation for the regression of the electronic energy.

A particular aspect of some electronic properties, such as the electron density [$\rho(\boldsymbol{r})$], is that they are local and depend significantly only on their immediate chemical environment.[253,254] Consequently, the representation of the chemical information to target these "nearsighted" properties benefits to be local as well.[255] In practice, even global extensive properties (such as the total electronic energy) can be partitioned into local or atom-based contributions and targeted with local representations. This procedure makes the regression scalable for large systems since local atom-centered representations can be transferred from smaller fragments (see, for instance, Chapters 5 and 6).

The Smooth Overlap of Atomic Positions (*SOAP*) is a local similarity measure between two atomic environments that allows bypassing altogether the use of a representation.[59] The main concept of SOAP is to compute the similarity (S) between neighboring environments as the inner product (overlap) of two atom-centered densities:

$$S(\rho,\rho') = \int d\boldsymbol{r}\,\rho(\boldsymbol{r})\rho'(\boldsymbol{r}) \tag{2.28}$$

where $\rho(\boldsymbol{r})$ are approximated by a sum of Gaussian functions centered on each atom of the neighborhood (χ),

$$\rho_\chi(\boldsymbol{r}) = \sum_{i\in\chi} \exp\left(\frac{(\boldsymbol{r}_i - \boldsymbol{r})^2}{2\sigma^2}\right) \tag{2.29}$$

The measure $S(\rho,\rho')$ is invariant over the permutation of atomic environments, but not over changes in their orientation. Nevertheless, the rotational invariance can easily be recovered by integrating over all possible orientations of one of the two environments:

$$\bar{K}(\rho,\rho') = \int d\hat{R} \left| \int d\boldsymbol{r}\,\rho(\boldsymbol{r})\rho'(\hat{R}\boldsymbol{r}) \right|^n \tag{2.30}$$

where $n \geq 2$ is an integer exponent needed to prevent the two integrals to be exchanged and thus to preserve the angular information. For most applications and to raise the sensitivity of the SOAP kernel to changes in atomic positions, it is beneficial to normalize the kernel and enhance its non-linearity by raising it to some power $\zeta \geq 2$,

$$K(\rho,\rho') = \left(\frac{\bar{K}(\rho,\rho')}{\sqrt{\bar{K}(\rho,\rho)\bar{K}(\rho',\rho')}} \right)^\zeta \tag{2.31}$$

For the applications relevant to this thesis, both exponents in Equations 2.30 and 2.31 are set to 2.

In practice, the evaluation of the integrals in Equation 2.30 is greatly simplified if the atom-centered densities [$\rho_\chi(\boldsymbol{r})$] are expressed in a basis composed of orthogonal radial functions and spherical harmonics for the angular part. In this way, it is possible to construct a SOAP representation in the form of a power spectrum, whose elements for each pair of atoms (Z_1, Z_2) are defined as

$$p(\boldsymbol{r})^{Z_1,Z_2}_{n,n',l} = \pi\sqrt{\frac{8}{2l+1}}\sum_m c^*(\boldsymbol{r})^{Z_1}_{n,l,m}c(\boldsymbol{r})^{Z_2}_{n',l,m} \tag{2.32}$$

with $c(\boldsymbol{r})^{Z}_{n,l,m}$ representing the projections of $\rho_\chi(\boldsymbol{r})$ onto the orthogonal basis set. Using the power spectrum representation, the SOAP similarity kernels take the form of a dot product.

$$K(\boldsymbol{p},\boldsymbol{p}') = \left(\frac{\boldsymbol{p}\cdot\boldsymbol{p}'}{\sqrt{(\boldsymbol{p}\cdot\boldsymbol{p})(\boldsymbol{p}'\cdot\boldsymbol{p}')}}\right)^{\zeta}. \tag{2.33}$$

2.4.2 Symmetry-adapted Gaussian process regression

The SOAP kernel, as defined in the previous section, is invariant under translation, rotation, and permutation of the environments and encodes all the fundamental symmetries to target a scalar molecular property. On the other hand, the formalism is not covariant with symmetry operations applied to any arbitrary tensor of an order higher than zero (*e.g.*, multipole moments, polarizabilities, stress tensor, *etc.*).[256,257] This poses no particular problem for small, rigid molecules (*e.g.*, CO_2), as it is possible to regress the properties in a local frame of reference where all the tensors would be aligned.[46,258] On the other hand, this limits tremendously the chemical diversity that can be targeted with SOAP-based GPR.

The generalization of the SOAP similarity kernel for the regression of any arbitrary-order tensor can be obtained by averaging over all symmetry operations $\hat{S}$:[47]

$$k_{\mu\nu}(\rho,\rho') = \int d\hat{S}_{\mu\nu}k(\rho,\hat{S}\rho') \tag{2.34}$$

where $\mu\nu$ are indices of two sets of axes. In particular, the requirement for tensor regression is that the kernel should be covariant with the group of all rotations about the origin of the 3D Euclidean space [SO(3) group]. According to SO(3) algebra, any Cartesian tensor can be decomposed in its spherical components as a linear combination of spherical harmonics.[259] In this basis, the transformation under rotation of any tensor is represented by the Wigner $\boldsymbol{D}^\lambda$

matrix.[260,261] Accordingly, the SOAP similarity kernel (Eq. 2.30) has to be modified as:

$$\bar{K}^{\lambda}(\rho,\rho') = \int d\hat{R}\boldsymbol{D}^{\lambda}(\hat{R})\left|\int d\boldsymbol{r}\rho(\boldsymbol{r})\rho'(\hat{R}\boldsymbol{r})\right|^{n}. \tag{2.35}$$

These symmetry-adapted kernels encode the correct geometrical transformations of tensors of any arbitrary rank and do not require alignment of the molecules to a fixed reference frame. Using these matrices as a measure of the covariance between different atomic environments within a symmetry-adapted GPR framework (*SA-GPR*) allows the scalable and transferable regression of any molecular property.

3 Balancing DFT Interaction Energies in Radical Cation Dimers

This chapter is based on the following publication:

A. Fabrizio, R. Petraglia, C. Corminboeuf, Balancing DFT Interaction Energies in Charged Dimers Precursors to Organic Semiconductors, *J. Chem. Theory. Comput.*, submitted. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.11309480.v1.

3.1 Introduction

The necessity of characterizing and validating the robustness of density functional approximations (DFAs) has motivated the construction of a variety of benchmark databases (*e.g.,* Ref. 262 and datasets therein) that target either specific chemical properties or specific classes of compounds. These collections of highly accurate data have aided in the development of improved functionals and dispersion corrected schemes that have increased the applicability and robustness of the Kohn-Sham density functional theory (KS-DFT) framework.[11]

In the last two decades, several databases focusing on intermolecular interactions have flourished.[263] Besides the fact that non-covalent interactions are ubiquitous and crucial for understanding molecular structures and properties,[134] the interest in non-covalent interactions is also methodological. In fact, accurately describing intermolecular interactions, especially Van der Waals forces, has been a longstanding challenge within Kohn-Sham density functional theory.[134,264,265]

In practice, one can categorize existing databases of intermolecular interaction energies into three distinct classes. The first includes datasets that specifically target non-covalent interactions, which includes Hobza's popular S22[266] along with its corrections[267,268] and extensions,[269,270] as well as the S66[271] set, Grimme's S12L,[272] Sherrill's NBC10[273,274], and the SSI databases.[275] A second class consists of databases assessing interaction energies along with other thermochemical and kinetics properties such as GMTKN30[276,277] along with its extension,[262] and Zhao and Truhlar's NCIE53.[140] Finally, a third category is oriented towards machine-learning and data-driven (combinatorial) applications and differs from the others by the exceptional size of the dataset (*i.e.*, thousands of entries). This class includes the databases of Friesner,[278] Shaw,[279] Head-Gordon[280] and the recent BFDb from Sherril group.[61] Overall, the large majority of these databases focuses on the intermolecular interactions between two closed-shell molecules constituting a neutral dimer, and only rarely between charged open-shell molecules.

Yet, charged open-shell dimers are equally relevant for both the methodological and application purpose. The smallest functional units of organic electronic materials (*e.g.*, organic photovoltaics,[281] organic field-effect transistors[282] and organic light-emitting diodes[283]) are, for instance, composed of π-dimer radical cations.[283–285] The Orel26rad[60] database, introduced in 2012 by one of us, illustrates the rather poor accuracy of functionals augmented by an atom-pairwise dispersion energy correction. This poor performance arises from the combined effect of the incomplete description of London dispersion interactions and the delocalization error, which can be pinpointed as the source of dramatic failures not only in radical cations dimers[60,286], but also in halogen,[287] pnicogen[288] and chalcogen bonds,[289] as well as solvated ions.[290,291] In particular, dispersion-corrected semilocal and hybrid functionals tend to strongly overbind radical cation dimers at their equilibrium structures and exhibit incorrect asymptotic behavior. On the other hand, long-range corrected exchange functionals improve the description of the asymptotic region, but severely underestimate the interaction energies at equilibrium. In general, these functionals perform not better than dispersion-corrected HF, which indicates that the poor performance of DFAs is rooted in a lack of medium-short range correlation contributions.[60,292,293]

While Orel26rad provides a useful set of π-dimer radical cations motifs and interaction energies, both the size and the complexity of the database remain limited. In practice, organic electronics involve large molecules that feature a variety of possible packing motifs.[294–296] This raises a question regarding extending the conclusions drawn from the Orel26rad compounds to larger and more realistic dimer motifs. In particular, the ability of the functionals to balance delocalization error and London dispersion interactions must be addressed for larger dimers. Indeed, both these effects will grow with the system size[128,272] but the exchange and dispersion energies will decay at a different pace.

Here, we build and analyze an extension to Orel26rad, which includes 9 large dimers used in practice as organic semiconductors (CryOrel: **Cry**stal of **Or**ganic **El**ectronics). The CryOrel9 dimers are representative of three different crystal arrangements : brickwork, herringbone and columnar-lamellar packing.[297] CryOrel9 is aimed at testing the accuracy of standard DFAs and wavefunction-based methods (such as USAPT0, MP2 and RPA) beyond the Orel26rad model systems, as well as at identifying those chemical situations and crystal motifs that are prone to larger errors.

To analyze these large dimers, we expand the original scope of functionals tested on Orel26rad and parametrize a variant of the ωB97X functional[298] jointly-fitted with the density-dependent dispersion correction (dDsC).[169–171] This variant is compared with its parent density functionals, ωB97X-D3,[299] ωB97X-V,[300] ωB97M-V,[280] and ωM06-D3[299], which clarifies whether these approaches are suitable to achieve an accurate description of organic electronics units both in their charged and neutral states.

3.2 Methods and Computational Details

3.2.1 The CryOrel9 set

The CryOrel9 dataset is aimed at overcoming specific limitations of OREL26rad, while retaining a focus on π-dimer radical cations that are relevant for the field of organic electronics. OREL26rad, as well as its neutral counterpart Pi29n, includes the simplest model systems of charge carriers for organic electronics arranged at their gas-phase minima. In contrast to the OREL26rad model systems, the nine CryOrel9 compounds are precursors of known semiconductors that are up to five times larger and arranged according to their experimentally determined crystal arrangement (Figure 3.1). While the herringbone packing is the most common, the brickwork arrangement is typical of chemical situations, where a π-conjugated core is functionalized with sterically demanding lateral substituents, such as in TIPS-pentacene.[301] This class of molecules with layered packing motifs generally exhibits the most efficient transport properties.[285] The last columnar/lamellar packing is typical of disk-shaped molecules such as hexabenzocoronene, which forms liquid-crystal phases upon substitution with floppy peripheral groups.[302]

Following the same protocol that was used for Orel26rad,[60] monomer geometries extracted from the X-Ray data were optimized at B3LYP[82,303,304]/6-31G* in Gaussian 16.[305] The radical cation dimers are constructed by assembling a neutral and a radical cation monomer for each compound without further relaxation. To preserve the crystal packing symmetries, the center of mass (C.O.M.) distances and the tilt-angles between the monomers are fixed to experimentally determined values. X-Ray data were taken either from original literature (ETTDM-TTF,[306] BDT[307]) or the Cambridge Structural Database[308] (DITT,[309] FPP-DTT,[310] BBBT,[311] DBT-Sulfone,[312] QTH,[313] DBT,[314] BTTT[315]). A dataset containing all relevant files will be made available upon publication in the Materials Cloud public repository.

3.2.2 Benchmark level

Interaction energies of the nine dimers were computed at estimated DLPNO-CCSD(T)[316,317] /CBS as implemented in the ORCA code.[318] The success of DLPNO-CCSD(T) for benchmarking large chemical systems is undoubtedly due to its favorable scaling, as well as its established accuracy.[317,319,320] Nevertheless, correlation energies converge slowly with respect to the basis set size and extrapolations to complete basis set (CBS) are needed. Following the procedure proposed by Neese and coworkers,[321] the interaction energies were computed after a two-point extrapolation using Dunning's cc-pVDZ and cc-pVTZ basis sets.[322]

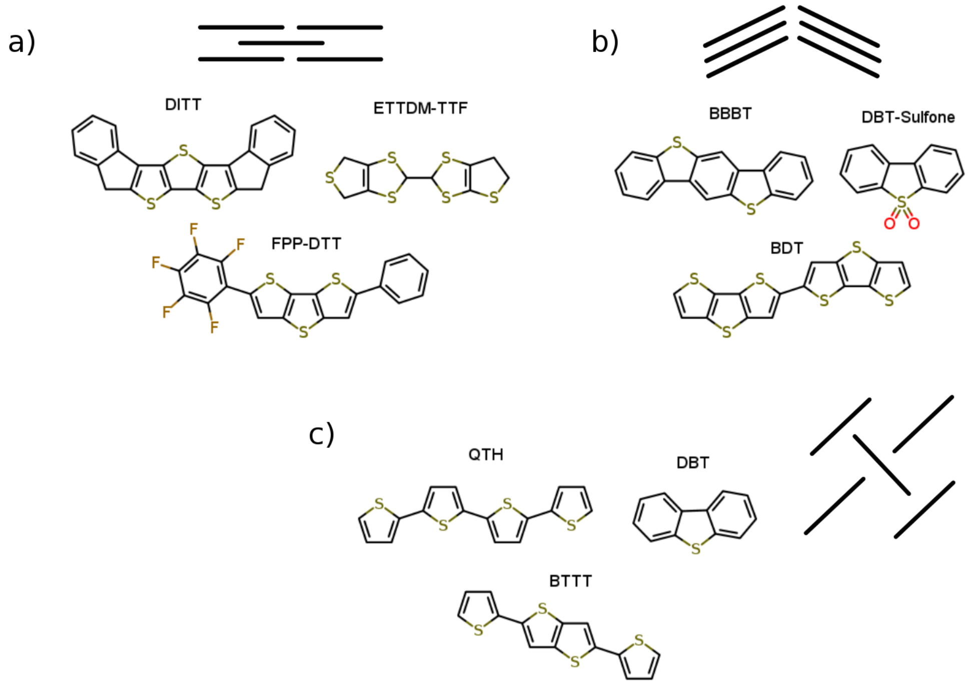


Figure 3.1 – The CryOrel9 set classified by the characteristic crystal packing of each compound a) 2D brickwork, b) columnar-lamellar and c) herringbone packing. The meaning of abbreviations in the Figure is detailed in Appendix A (Table A.2).

The Boys-Bernardi scheme was applied to all computations to correct for basis set superposition error (BSSE).[323] To avoid errors stemming from spin-contamination, computations were performed at the ROHF-DLPNO-CCSD(T) level.[324]

3.2.3 Density functionals and wavefunction based computations

Two main aspects of the density functionals have been tested on the CryOrel9 database: the influence of the fraction or range of exact exchange and the sensitivity of the result to different dispersion correction schemes. Additional details on the functional tested and their performance can be found in the Appendix A.

DFT computations were performed using the def2-TZVP basis set[325] and a fine grid for the meta-GGA and meta-hybrid functionals. Computations with the GGA, global hybrids, TPSS[326] and ωB97X-D were performed with GAMESS-US;[327,328] the Minnesota functionals were tested in Gaussian 16;[305] ωB97X-D3, ωB97X-V, ωB97M-V and ωM06-D3 were computed using QChem.[329]

Along with these common DFAs and dispersion corrections, three wavefunction based methods (*i.e.*, RPA, MP2, and U-SAPT0) were also tested. In principles, RPA can be also thought of as an extension of the density functional test set that seamlessly includes non-local correlation effects and exact HF-exchange.[330] MP2 and U-SAPT0 computations were performed in combination with the 6-31G*(0.25) and the jun-cc-pVDZ basis sets, respectively.

Random phase approximation[331,332] (RPA) computations used the resolution of identity and the frozen core approximation in Turbomole 7.1.[333] The self-consistent Kohn-Sham orbitals were obtained from previously converged PBE[334,335] computations. The complete basis set results (RPA/CBS) were computed using three-point extrapolation, following established protocols.[336] MP2/6-31G*(0.25)[337] computations were performed in Molpro 2015,[338] using the resolution of identity approximation. Unrestricted symmetry-adapted perturbation theory (USAPT0[339]) was performed as implemented in the Psi4 software package,[340] along with the suggested jun-cc-pVDZ basis set.[341]

3.2.4 Fitting and validation of ωB97X-dDsC

ωB97X-dDsC belongs to a larger family of dispersion-corrected range-separated hybrid density functionals derived from Head-Gordon and Chai's ωB97X. Their general structure relies upon the B97-type expansion of inhomogeneity-correction factors (ICFs),[342] combined with a fixed fraction of HF-exchange that gradually increases up to 100% in the long-range. Each functional declines this common paradigm as well as the parametrization in a different way. For instance, ωB97M-V introduces an additional dependence on the kinetic energy density and results from the systematic generation and testing of a combinatorial library containing approximately 10^{10} candidate functional forms. Yet, a common and very relevant feature in the present context is that all the functionals in the series are jointly-fit with a dispersion correction. Chai pursued the same strategy with ωM06-D3, combining range-separation and joint-fitting of the dispersion correction on top of the M06 meta-hybrid functional.

The choice of the dispersion scheme provides a further classification. ωB97X-D and ωB97X-D3 are based on *ad-hoc* atom pairwise corrections that account for only one class of non-additive effects (type-A, following the Dobson's classification[343]). It is worthwhile noting that non-additive effect beyond the pairwise approximation (type-B) can be accounted for through the addition of a three-body term (Axilrod-Teller-Muto) to DFT-D[144,344] or to infinite-order as in the many-body-dispersion scheme (MBD) of Tkatchenko and coworkers,[145] as well as in RPA and CCSD.[330] Theses effects are crucial for large systems, as demonstrated on nanoscale materials[345] and large fullerenes.[346] Nevertheless, the size of the dimers studied here is significantly smaller than these examples and pairwise additivity remains a practical choice. In fact, the relative importance of many-body contributions with respect to higher-order pairwise terms (C_8 and C_{10}) is still a subject of debate.[146,272,347,348] In particular, it has been demonstrated that, even for small dimers, C_8 and C_{10} capture nearly half the dispersion energy[159] and that, when omitted, the leading C_6 term mimics their role and results in a systematic overbinding of dispersion interactions.[146,348] In this respect, the dispersion correction of ωB97X-dDsC includes all dispersion pairwise coefficients up to third-order (C_6, C_8 and C_{10}). The more recent variants tested hereafter, ωB97X-V and ωB97M-V, include an explicitly non-local correlation term rooted in VV10.[137,343]

Akin to ωB97X-D and ωB97X-D3, ωB97X-dDsC is tuned by 15 adjustable parameters (5 for the exchange gradient-correction factors, 5 for the same-spin correlation and 5 for the opposite-spin correlation). Four additional parameters determine the range-separation, control the fixed fraction of exact exchange at short-range and tune the damping function of the dispersion correction.

As for the other members of the ωB97X series, ωB97X-dDsC obeys the uniform electron gas limit. A constraint that fixes 3 functional parameters and allows the remaining 16 to vary freely. The task of simultaneously optimizing these parameters was tackled with a time- and memory-efficient algorithm divided into two recursive sub-procedures, which rely on partial parameter optimization using frozen densities. Given the high dimensionality of the optimization problem, the dependence on initial conditions of the parameters was evaluated by perturbing the optimized set. The parameters of the proposed functional were found to be stable to a 10% perturbation. Figure 3.2 is a diagrammatic illustration of the sequential procedure used in the determination of the 19 adjustable parameters that tune ωB97X-dDsC.

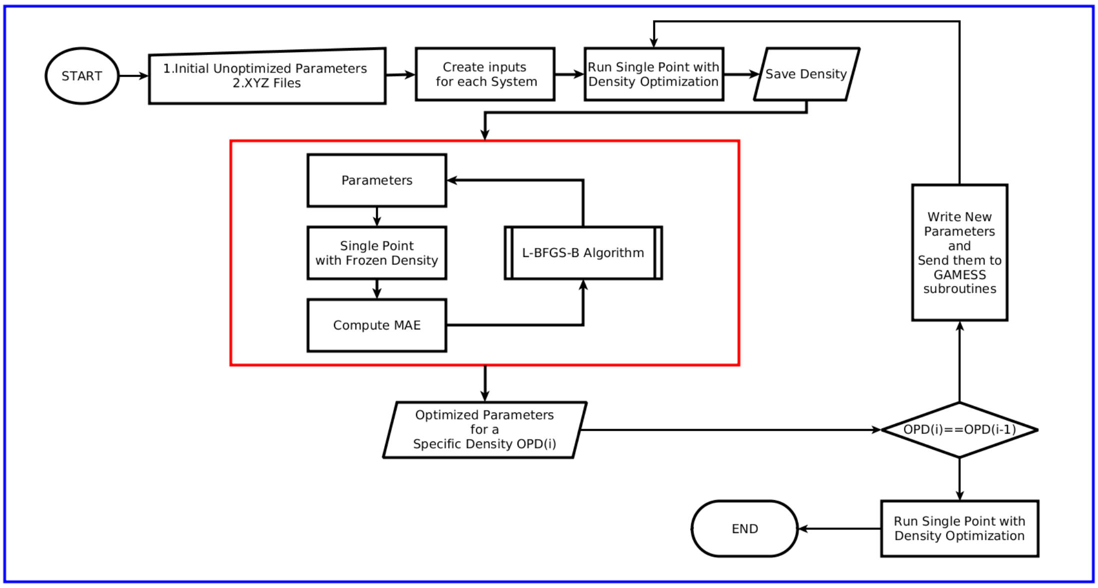

START
1.Initial Unoptimized Parameters
2.XYZ Files
Create inputs for each System
Run Single Point with Density Optimization
Save Density
Parameters
Single Point with Frozen Density
L-BFGS-B Algorithm
Compute MAE
Write New Parameters and Send them to GAMESS subroutines
Optimized Parameters for a Specific Density OPD(i)
OPD(i)==OPD(i-1)
END
Run Single Point with Density Optimization

Figure 3.2 – Main algorithm used for the optimization.

The main optimization algorithm is divided into two recursive procedures, delimited in Figure 3.2 by a blue and a red box. The first one (in blue) is responsible for data generation and storage, input handling and decision making, while the nested processes (in red) perform the actual optimization.

The algorithm requires the user to input the molecular geometry of each molecule of the training set (xyz format), as well as to provide the initial guess for the parameters to optimize. Those data are used to generate input files compatible with a modified version of the GAMESS program package[327,328] and then to run a single point computation where the electronic densities are optimized self-consistently. The relaxed densities and the effective atomic polarizabilities are then stored as external files in the memory (RAM). Using the data stored in the RAM, the core optimization process computes single point energies and returns the mean absolute error (MAE) of the training set relative to a specific array of parameters. The MAE is then minimized using a constrained version of the Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS-B)[349], in order to find the optimal parameters for a specific density (OPD(i)). At this point, the algorithm is able to compare the initial and the updated parameters and decide to either conclude the optimization or to iterate the cycle. The convergence criterion is met when the array of parameters remain unchanged after the optimization process.

The original GAMESS code was adapted to the necessity of reading functional parameters from an external input and storing important data, densities and polarizabilities in external unformatted files. A renewed version of the subroutines for the computation of the dDsC dispersion correction was also implemented in the final operational version. The optimization core was designed to use only 14 subroutines of the entire GAMESS program package in order to return the single point energy for each chemical system.

Both parameter training and validation of the functional were performed using a modified version of the GAMESS-US program package.[327,328] The full list of optimized parameters is available in Appendix A of the original paper (Table A.1). During the functional parametrization, numerical integrations were performed on the fast SG-1 grid[226] (50/194), while for validation a finer Euler-Maclaurin-Lebedev grid was used (75/302). The parametrization and validation of the functional were performed using the def2-TZVPD basis set. Since no significant dependence on diffuse functions was found, all other computations were performed with the def2-TZVP basis set.

The 19 adjustable parameters of ωB97X-dDsC have been optimized to reduce the mean absolute error (MAE) on the same training set originally as used for ωB97X-D and in large part for ωB97X-D3. As reported in Figure 3.3, this set includes atomization energies (G2), ionization potentials (IP), electron affinities (EA), proton affinities (PA), hydrogen transfer barrier heights (HTBH), non-hydrogen transfer barrier heights (NHTBH), and non-covalent interactions (S22).

The general performance of the trained ωB97X-dDsC is similar to that of ωB97X-D, with only a slight overall improvement of about 0.2 $\text{kcal}\cdot\text{mol}^{-1}$ over the whole set. This result is not surprising considering that ωB97X-dDsC and ωB97X-D share the same exchange-correlation functional form. Exceptions to these similar performances include proton affinities (PA), as well as forward and reverse barrier heights in the HTBH set for which the MAE of ωB97X-dDsC decreases by about 0.7 $\text{kcal}\cdot\text{mol}^{-1}$ with respect to ωB97X-D.

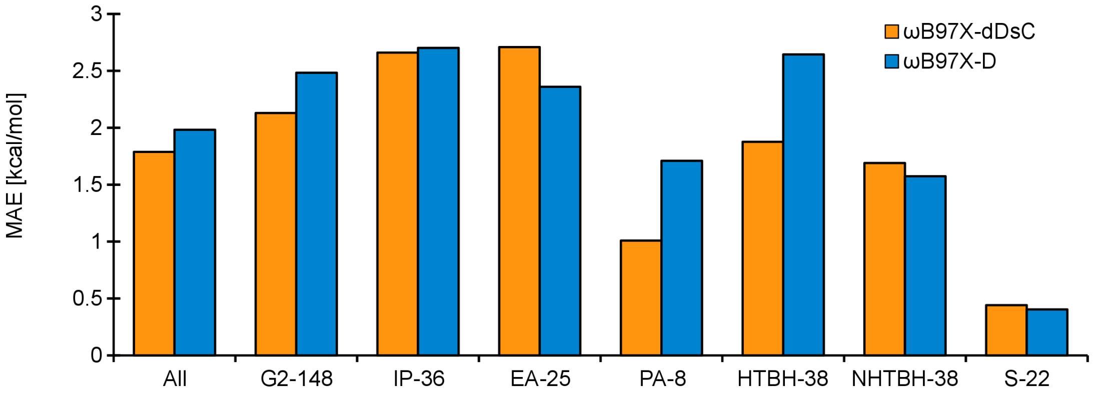


Figure 3.3 – Mean absolute error (MAE) of ωB97X-dDsC on the training set. ωB97X-D is shown for the sake of comparison.

3.3 Results and discussion

Having demonstrated the accuracy of ωB97X-dDsC on the training set, we now compare its performance with that of standard dispersion-corrected density functional approximations, including the ωB97X variants and ωM06-D3 on the Orel26rad and Pi29n datasets. The averaged performance of the functionals is reported in Figure 3.4 using two statistical metrics to probe their absolute (MAE: mean absolute error) and their relative deviation (MAPE: mean absolute percentage error).

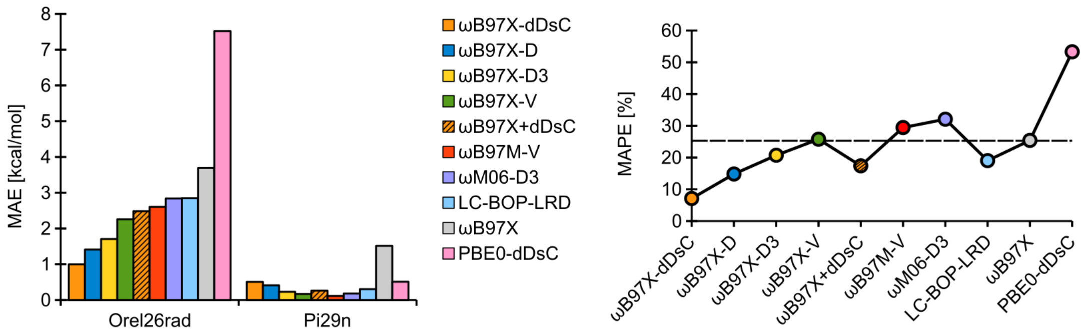


Figure 3.4 – (*left*) MAE of illustrative functionals on the radical cation dimers of Orel26rad and the corresponding neutral compounds of Pi29n.(*right*) MAPE of the same functionals on the Orel26rad dataset. The horizontal line represents the MAPE of ωB97X.

Overall, impressive results are achieved for the neutral, closed-shell dimers in Pi29n, for which each of the dispersion corrected functional variants tested herein achieved errors lower than 1 kcal · mol^{-1}. In particular, the best performing ωB97M-V (MAE of 0.1 kcal · mol^{-1}) matches its MAE on the neutral van der Waals dimers in the original training set.[280]

The mean absolute error on the radical cation dimers, in contrast, varies substantially among the ωB97-series, ranging from the 1 kcal$\cdot$mol^{-1} MAE for ωB97X-dDsC to 2.84 kcal$\cdot$mol^{-1} for ωM06-D3. The three best-performing variants: ωB97X-dDsC, ωB97X-D, and ωB97X-D3 share a common design principle and are based on a uniformly truncated and high-order (quartic) polynomial series of inhomogeneity correction factors (ICF).[155,299] This principle has been abandoned in recent variants starting from ωB97X-V, as individual determination of the truncation orders for each component of the functional allowed for a reduction in the number of fitted parameters with only small variations in the overall accuracy on the original training set.[300] The slight accuracy loss on Orel26rad (The MAEs of the meta-GGA variants, ωB97M-V reach 3.00 kcal$\cdot$mol^{-1}) suggests that reducing the flexibility of the core functional (*i.e.* the non-dispersion term) might alter the robustness.

The relative metric, mean absolute percentage error (MAPE), provides further information on the robustness. While the MAEs probe the magnitude of the error, the MAPEs indicate whether these deviations are proportional to the magnitude of the interaction energy. The MAEs and MAPEs on Orel26rad follow the same trend except for ωB97X, ωB97X+dDsC and LC-BOP-LRD. The MAE of ωB97X is one of the largest but its MAPE is lower than ωM06-D3 and ωB97M-V. This mismatch implies that even if ωB97X leads to large errors, its deviation is larger for systems with the highest interaction energy (owing to the missing dispersion correction term). The errors of the dispersion-corrected variants characterized by a MAPE higher than ωB97X (*i.e.*, ωB97M-V and ωM06-D3) correlate less with the magnitude of the interaction energies and are, in this respect, less systematic.

The advantage of the joint-fitting of the functional and dispersion correction (*i.e.*, ωB97X-dDsC) is assessed in Figure 3.4 by comparison with the original ωB97X, where the dDsC dispersion correction has been fitted *a posteriori* without modifying the functional parametrization (noted as ωB97X+dDsC). ωB97X+dDsC performs well both radical cationic dimers and the neutral Pi29n complexes. For Orel26rad, ωB97X+dDsC has a MAE similar to the most recent ω-variant, but a much smaller MAPE. Hence, modification of the ωB97X core is at the origin of the poorer scaling with the interaction energy magnitude in ωB97X-V and related forms. Comparing ωB97X+dDsC and ωB97X-dDsC highlights the advantage of jointly fitting the functional and the dispersion correction parameter, especially for Orel26rad (where the error magnitude is cut in half), at the condition of keeping the same flexibility in the core as the parent ωB97X. Analyzing the two damping function parameters for the two approaches reveals a two-fold increase in both parameters a_0 and b_0, which correlates with the doubling of the parameters tuning the leading orders of the opposite-spin correlation term in ωB97X-dDsC[171] (Appendix A).

Figure 3.4 reports energies associated with equilibrium geometries but the ability to properly describe the interactions of radical charged dimers out-of-equilibrium is as relevant. In line with our work on the OREL26rad database,[60] we compute the interaction energy profiles of (furan)$_2^{\cdot+}$ and (thiophene)$_2^{\cdot+}$ with the ωB97X-series and ωM06-D3 (Figure 3.5). The challenging nature of these two energy profiles is evident when compared with the PBE0-dDsC and LC-

BOP-LRD profiles.

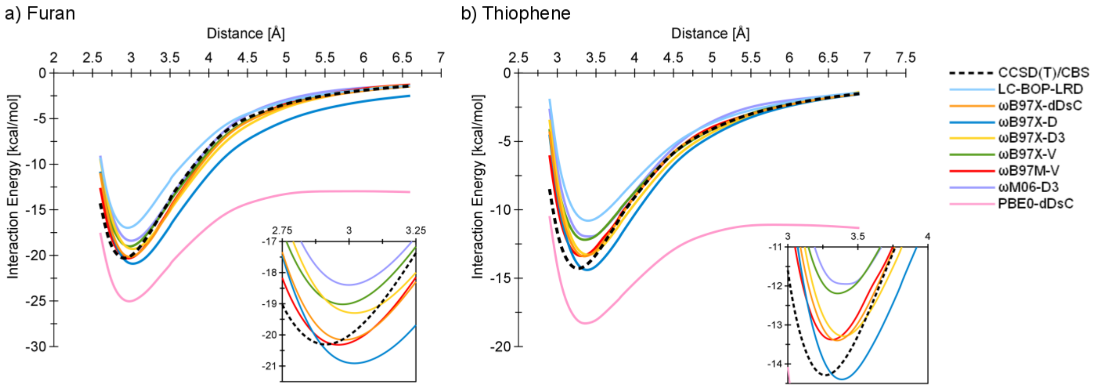


Figure 3.5 – Interaction energy profiles for the radical cation dimers of a) furan and b) thiophene. Insets zoom into the equilibrium region. CCSD(T)/CBS, LC-BOP-LRD and PBE0-dDsC values are taken from Ref. 60.

All the dispersion-corrected range-separated DFAs tested retrieve the CCSD(T)/CBS limit in the asymptotic region with the exception of ωB97X-D for the furan dimer. The deviation suggests that a ω value higher or equal to 0.25 is necessary to recover the correct asymptotic behavior in radical cation dimers.

More variations are observed closer to the equilibrium region as illustrated by the too repulsive ωB97X-V, and ωM06-D3 that contrast with the best performing ωB97X-dDsC and ωB97M-V.

The common features leading to the best performance on OREL26rad rely upon a uniformly truncated and highly flexible exchange-correlation core fitted jointly with an atom-pairwise damped dispersion correction. Despite their higher MAEs, the more recent ω-variants are highly accurate on specific systems, as exemplified by the ωB97M-V profiles. This variability in performance suggests that their robustness across size and molecular arrangement is limited as the errors are more system dependent. These observations further motivate the construction of a more challenging set of radical cation dimers.

3.3.1 CryOrel9 dataset

The mean absolute errors of three wavefunction-based methods (*i.e.*, RPA/CBS, MP2/6-31G*(0.25), and U-SAPT0/jun-cc-pVDZ) are evaluated on the CryOrel9 dataset with respect to the DLPNO-CCSD(T) reference data (see Figure 3.6-a).

The accuracy of RPA on neutral van der Waals assemblies has been assessed extensively for both molecules[350–354] and solid-state systems.[355–357] With an overall MAE slightly above 0.5 $\text{kcal} \cdot \text{mol}^{-1}$ for CryOrel9, RPA/CBS appears as a reliable benchmark for charged assemblies of radical cations. Given the relatively small average difference between DLPNO-CCSD(T) and RPA, it is legitimate to wonder which of the two methods would be the preferable ref-

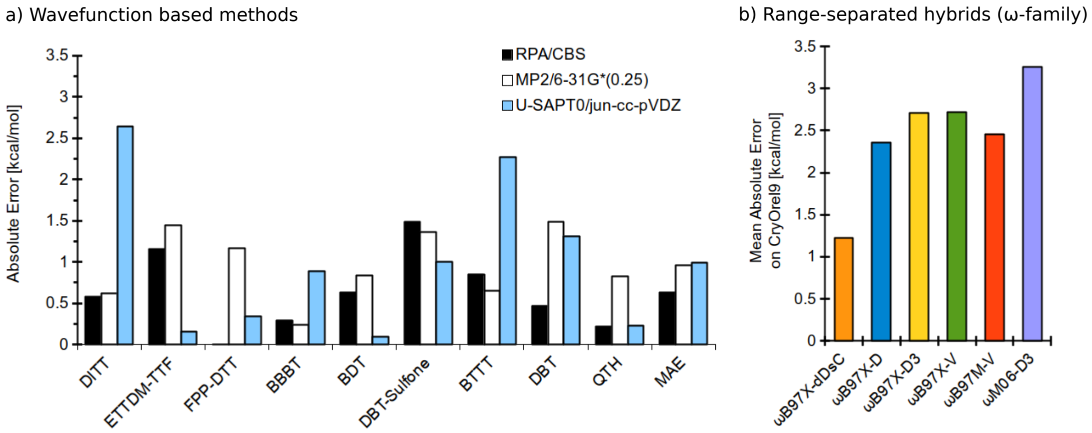


Figure 3.6 – Mean absolute error of a) the tested wavefunction based methods and b) the range-separated functionals of the ω-family on the CryOrel9 with respect to DLPNO-CCSD(T)/CBS reference.

erence level. In this respect, it is worth noting that the RPA ground-state correlation energy is only equivalent to a ring-diagrams simplification of the coupled-cluster doubles (rCCD) equations,[358] while DLPNO-CCSD(T) contains all the diagrams of the CCSD(T) method, with a restriction of the excitation only to the most relevant virtual orbital subspace.[316] On this basis, DLPNO-CCSD(T) is retained as the preferred reference method.

The CryOrel9 MAE of the less computationally demanding MP2/6-31G*(0.25), is rather low (*i.e.*, about 1.00 kcal $\cdot$ mol^{-1}) but twice that of OREL26rad,[60] showing the limit of exploiting the error cancellation between MP2 and the modified basis set. Also relying on error cancellation,[341] the performance of U-SAPT0/jun-cc-pVDZ is here similar to MP2/6-31G*(0.25).

In addition to its reasonable performance on CryOrel9, U-SAPT0 provides a chemical rationalization of the individual energetic contributions to the total interaction energy. Comparing OREL26rad and CryOrel9 highlights that both the absolute and the relative contribution of London dispersion associated with CryOrel9 double at the detriment of induction (Figure 3.7, pie charts and bars). For this set, the sum of two contributions, London dispersion and the exchange, represent 75% of the total interaction energy. This percentage is significantly higher than Orel26rad, for which the four contributions are balanced.

The increase of the absolute and relative contributions of London dispersion interactions is readily explained by the dimer sizes (more atom pairs and larger polarizability). The lowering of the induction term is also related to the larger size of the cationic monomer, as the density of the hole (and the strength of the associated electrostatic field) is distributed among a larger number of atomic centers. Finally, the electrostatic interactions do not change significantly as expected from the total charge that remains identical between the two sets.

Even though the relative contributions to the interaction energy of all the crystal motifs in CryOrel9 are similar, the absolute contributions vary greatly with the relative packing

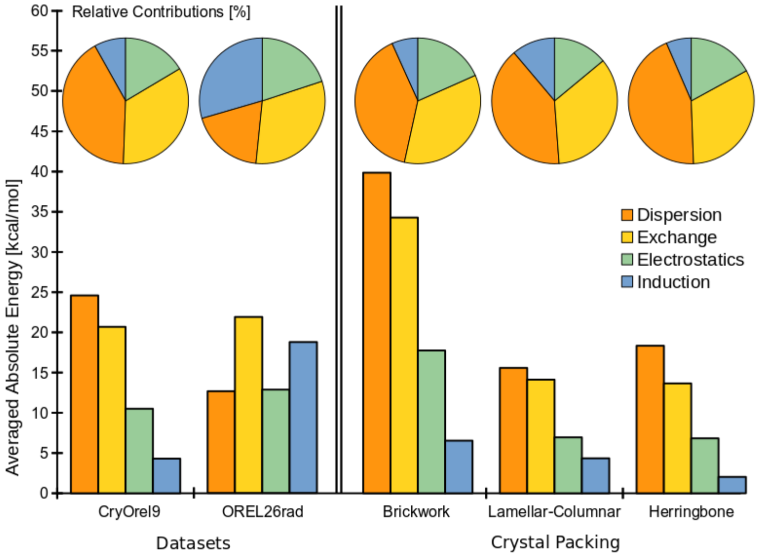


Figure 3.7 – Relative and absolute values of U-SAPT0 contributions averaged over CryOrel9 and Orel26rad (left) and divided per type of preferred crystal arrangement in CryOrel9 (right).

orientation. The U-SAPT0 analysis of each type of supramolecular arrangement identifies the brickwork packing as quantitatively different from the other motifs as it exhibits stronger interactions (Figure 3.7). In line with previously published SAPT analyses of π-stacked neutral dimers,[359,360] the larger magnitude of the London dispersion and exchange component in the brickwork motif is justified by the shorter distances and larger orbital overlap between the monomers with respect to other packing motifs.

In light of these considerations, it is not surprising that the overall functional performance on CryOrel9 (see Figure S3) is mainly dictated by their ability to accurately describe systems dominated by pronounced London dispersion interactions and large exchange contributions. Top panels of Figure 3.8 reports the correlation between the absolute error in the interaction energy for a selection of dispersion-corrected range-separated functionals and the absolute U-SAPT0 exchange and dispersion contributions.

Although the general accuracy fluctuates from dimer to dimer, none of the functionals has difficulty with the interaction energy of the leftmost dimer (BBBT), characterized by the lowest absolute dispersion and exchange contribution. The most problematic dimers belong to the brickwork motif (DITT and FPP-DTT) or correspond to the smaller stacked dimers from the columnar category (DBT-Sulfone). The common characteristic of these sensitive systems is evident when regrouping them according to the angle between their average monomer planes (tilt angle). This leads to two categories: stacked (angle $< 10°$) and tilted (angle $> 45°$) in the bottom panel of Figure 3.8. For each dimer, we report the standard deviation of the absolute error among the functionals tested. This variation is up to 5 times larger for the stacked dimers. The exception is ETTDM-TTF that has a small tilt angle but is highly asymmetric, as only one of its monomers is planar. By construction, all the dimers in CryOrel9

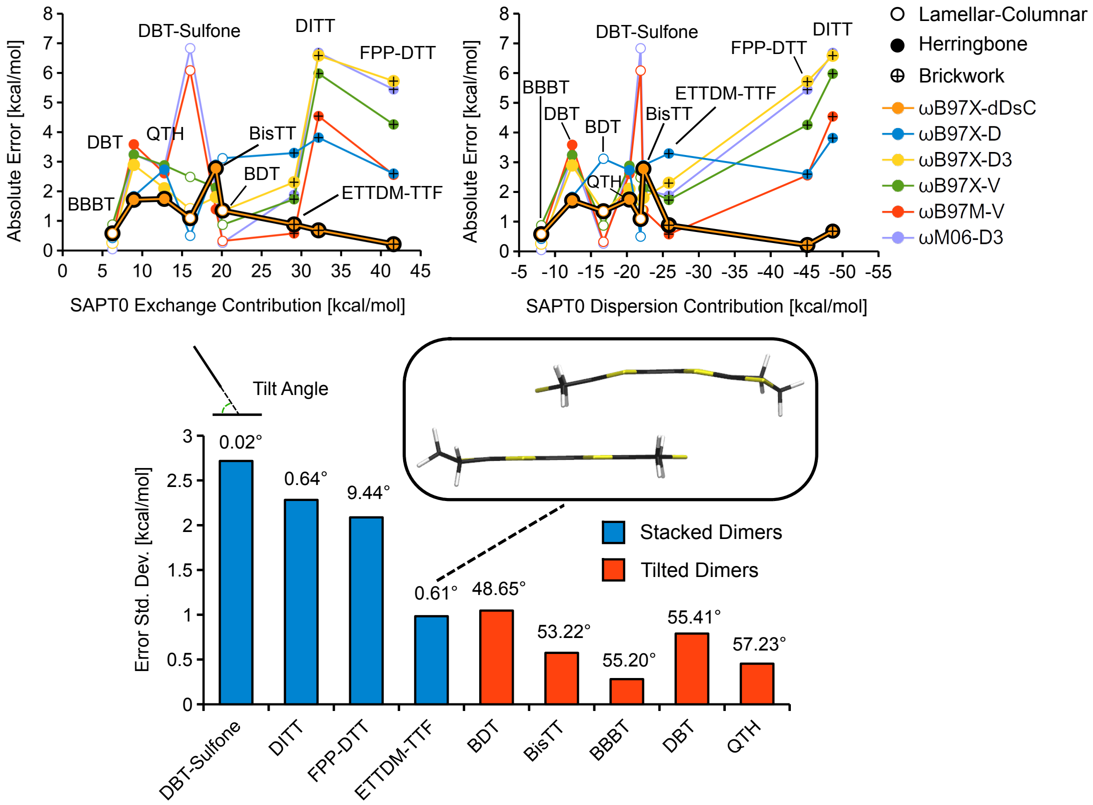


Figure 3.8 – (*top*) Correlation between absolute errors of the functionals with respect to DLPNO-CCSD(T)/CBS and the U-SAPT0 (*top left*) exchange and (*top right*) dispersion contributions. Each point represents one dimer of the CryOrel9 dataset. Color code represents different density functionals, while packing motifs are indicated with different symbols. (*bottom*) Spread of the error among the ω-family (standard deviation) for each dimer in CryOrel9. The color code highlights the classification on the basis of the tilt angle, reported on top of each histogram. The inset shows an orthographic view of the ETTDM-TTF dimer.

are built by assembling an optimized neutral and a radical cationic monomer at the crystal position without further relaxation (Section 2.1). However, the localization of the spin density on the monomer cation is less pronounced in the planar stacked dimers, and thus much more dependent upon the chosen functional. DBT-sulfone dimer represents an extreme case for which the most recent functional variants (*e.g.*, ωB97XM-V) place the spin density on the opposite "neutral" monomer (see spin-densities in Appendix A) causing the largest errors.

The interaction energy profiles of three dimers representative of each packing motif (Figure 3.9-a,b,c) completed by the ETTDM-TTF asymmetric example are shown in Figure 3.9-d. Generally, all the functionals retrieve the same asymptotic regime as the DLPNO-CCSD(T) reference but some of them suffer from instabilities in the form of a wiggling with increasing the intermonomer distance (see ωB97X-D3, ωB97M-V and ωM06-D3 in the inset of Figure 3.9-d) or/and an underestimation at the medium range (ωM06-D3). In the equilibrium region, the quality of the ωB97X-D profiles varies greatly depending on the dimers: BDT, ETTDM-TTF strongly underbind, whereas the profile of the herringbone DBT is highly accurate. On the performance spectrum, ωB97M-V and ωM06-D3 lead to exactly the opposite trends as they are excellent for ETTDM-TTF and BDT but underbind the rest. Overall, the ωB97X-dDsC profiles appear to be most robust across these specific systems.

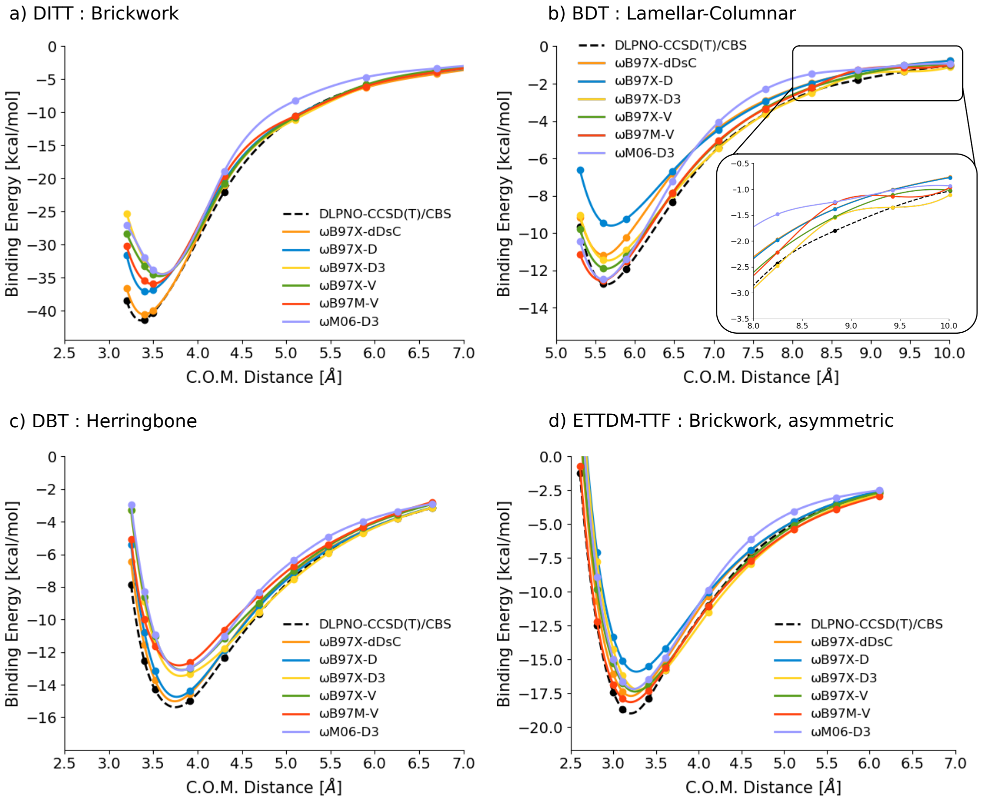


Figure 3.9 – Interaction energy profiles for the radical cation dimers of a) DITT, b) BDT, c) DBT and d) ETTDM-TTF.

In general, the evolution from the small model systems of OREL26rad to the relatively larger radical cationic dimers of CryOrel9 is associated with a redistribution of the relative importance of interaction energy contributions largely in favor of London dispersion interactions. This regime, where Pauli exchange and London dispersion contribute to more than 75% of the total interaction energies, does not significantly increase the range of errors of the tested approaches but highlights other trends. An important one lies in the overall deteriorating performances of ωB97X-D and ωB97X-D3 that are especially problematic with the Brickwork arrangements. The most recent variant ωB97M-V (and to a lesser extent ωB97X-V) built from statistical models essentially suffer from a singularity in the spin density of DFTB-Sulfones and from irregular MAEs ranging from very low for some systems but the largest for others. Although distinct by construction, ωM06-D3 also shows irregular errors but systematically underbind the profiles away from equilibrium. Amongst this functional series, ωB97X-dDsC performs best. While the reasons for this achievement are difficult to rationalize, they likely stem from the additional flexibility brought by the density-dependent double damping characteristic of dDsC.[169–171] In this dispersion correction, the argument of the universal Tang and Toennies damping function[152,153] is modified with a second damping term, which accounts for the regions of strong density-overlap (covalent-regime).[169,361] This combination provides better control of the medium-range correlation, which is generally underestimated in radical cation dimers.[60]

3.4 Conclusion

The quantum chemical description of radical cationic dimers that are key species in molecular organic materials is known to be challenging. The 2012 Orel26rad dataset[60] already illustrated the advantage of fitting jointly the dispersion correction with a long-range corrected exchange functional to provide both the proper dimer dissociation behavior and avoid underbinding at the equilibrium geometries. In this chapter, we construct CryOrel9, a more realistic set of radical cationic dimers that are extracted from distinct crystal arrangements including the brickwork and herringbone motifs.

We explore the performance of the ωB97X functional series on their interaction energy and dissociation profiles. To improve their description, we also parametrized a variant of ωB97X jointly-fitted with the dDsC density-dependent dispersion correction. RPA, USAPT0 and MP2/6-31G*(0.25) results are also provided and compared to the DLPNO-CCSD(T)/CBS reference.

ωB97X-dDsC is the most robust DFAs tested owing to the use of density overlaps in the damping function that control the damping in the medium range. The combinatorially-optimized variant ωB97M-V is highly accurate for some radical cationic dimers, but its robustness across chemical diversity and different spatial arrangements is not guaranteed. Alternatively, the parent functionals (*i.e.*, ωB97X-D and ωB97X-D3) are limited by the poorer description of the brickwork arrangement. In the prospect of overcoming the delicate interplay of errors that characterizes radical cation dimers, non-linear regression techniques could provide a solution. One foresees two strategies to achieve this goal. The first relies on the pragmatic application of system-dependent machine learning corrections to the interaction energy of approximate functionals with, for instance, Δ-ML. The second approach, which has been successfully applied to two-electrons, one-dimensional systems,[362] works on a more fundamental level and relies on the statistical learning of fully non-local exchange-correlation potentials. Given that the reliability of machine-learning models depends strongly on the accuracy of the underlying electronic structure methods used for their training, the approaches tested in this chapter as well as their comparison with the DLPNO-CCSD(T) reference appear essential. Such a data-driven effort would also involve the construction of much larger database in the spirit of BFDb[61] (or the database of Shaw and coworkers[279]) which was successfully used for training transferable machine learning model capable of predicting the electronic structure properties of large dimer systems.[34]

Finally, from a chemical perspective, U-SAPT0 on CryOrel9 showed that planar, π-stacked dimers are the most challenging systems, leading to large errors for most DFAs. This result is especially important considering the variety of molecular precursors of organic semiconductors that crystallize in a characteristic brickwork or columnar packing.

4 London dispersion and photochemical processes of molecular switches

This chapter is based on the following publication:

A. Fabrizio, C. Corminboeuf, How do London dispersion interactions impact the photochemical processes of molecular switches ?, *J. Phys. Chem. Lett.*, **2018**, *9*, 464-470.

Accurately describing ubiquitously present London dispersion interactions[363] has been a long-standing challenge for standard density functional approximations.[264,364,365] Today, several conceptually different approaches addressing this issue exist including the popular dispersion correction schemes DFT-D of Grimme and coworkers,[133,144,154,160] density-dependent dispersion corrections based on Becke-Johnson's exchange dipole moment (XDM[157–159,166–168] and dDsC[169–171]), Tkatchenko and Scheffler's TS[162] and MBD[145] methods, and the local-response dispersion (LRD) model by Sato and Nakai.[172,173] Aside from these *ad hoc* corrections, alternatives within the density functional theory (DFT) framework include the explicitly nonlocal density-based functionals (*e.g.*, vdW-DF[135,136], VV10[137]) and the effective one-electron potentials (*e.g.*, DCACP,[138,139] M06L[366]). However, each of these approaches was specifically derived to describe van der Waals (vdW) interactions in the ground state.[134] This raises a question regarding the adequacy of available methods to describe van der Waals interactions in excited states.

In practice, three scenarios exist where an incomplete treatment of London dispersion upon optical excitation could result in substantial errors (Figure 4.1). In the first case, the electronic transition occurs only on a localized portion of the molecule, while attractive long-range forces dominate the interactions between the remaining unexcited portions (Figure 4.1-a). This is, for instance, common in organic photovoltaics where chromophores are often functionalized with long alkyl side chains to improve solubility.[367] In the second case, the optical properties of a photoexcited molecule can change significantly when they interact with a ground-state molecule (Figure 4.1-b). In nature, this is seen in phenomenon such as co-pigmentation, which determines the stability and the modulation of flower and vegetable color,[368] as well as chemosensing[369] and (micro)solvated dyes. A third situation involves the interaction between two photoexcited molecules or fragments (Figure 4.1-c). Aromatic excimers typically belong to this category.[370–372]

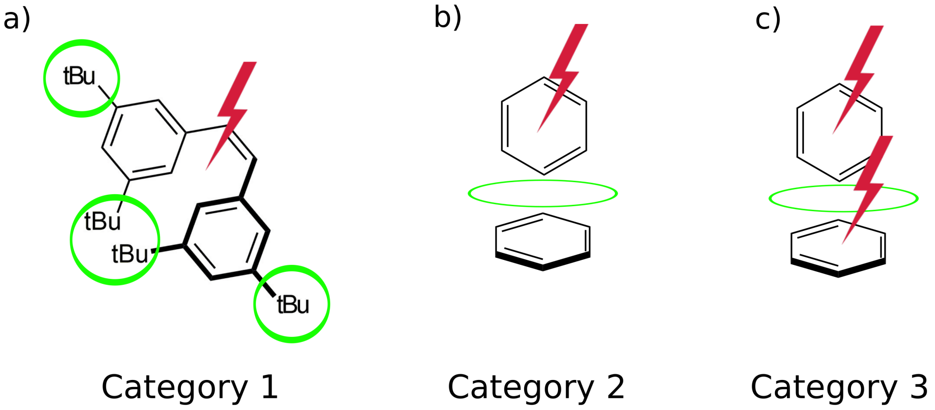


Figure 4.1 – Upon optical transition, London dispersion dominates the interaction between: a) unexcited fragments; b) a photoexcited and a ground state molecule; c) two photoexcited molecules.

Because of their inability to adapt to electronic transitions or dramatic structural changes, most of the dispersion correction schemes mentioned above would struggle for categories 2 and 3 (Figure 4.1-b and Figure 4.1-c), but might provide a qualitatively correct description of dispersion interactions for category 1 systems. The complexity for describing systems belonging to the latter two categories is only intensified by the inherent shortcomings associated, for instance, with functional choice in describing the excited states or with the adiabatic approximation in linear response time-dependent density functional theory (LR-TDDFT).[373]

Computational work addressing the role of vdW interactions in category 2 and 3 systems remains relatively scarce but efforts toward the development of sophisticated schemes, which capture London dispersion effects proper to excited states (*e.g.*, repulsive dispersion[374–376]) or which focus on the interaction of arbitrary state are currently on-going.[377–382] In general, these schemes require a fairly complex quantum electrodynamical treatment of the vdW-Casimir potential, so that their application have yet been restricted to nothing but the simplest chemical systems and perfectly homogeneous media. Other relevant examples include the analytical expression for the static polarizabilities of the s- and p-symmetric excited states of atoms[383] introduced by Adelman and Szabo but excited-state polarizabilities (and subsequent dispersion forces) require expensive computational techniques (such as the complex polarization propagator method) that are, generally, best for benchmarking.[384,385] Density matrix functional theory has also been extended[386] to treat vdW interactions for the prototypical two-electron triplet H_2 molecule, but the formalism cannot easily be adapted to treat larger systems. Within the context of LR-TDDFT, significant effort has been placed in achieving proper aromatic excimer energetics (category 3). The potential energy surfaces of aromatic excimers can be reasonable characterized at the TD-B3LYP level, even when the same functional fails dramatically for the ground state complex.[387] Subsequent work showed that the use of a dispersion corrected functional in the LR-TDDFT computation improved the dissociation energy curves of benzene/pyridine excimers.[388] A small number of studies evaluated the binding energies of excited complexes belonging to category 2 computed using either an atom pairwise correction[389,390] or the LRD model.[391] A noted improvement was generally observed, although the need for reparametrizing the correction or the functional was often advocated. Similar warnings were raised when using the innovative self-consistent field (SCF) algorithm alternative to TDDFT (i.e. the maximum overlap method).[392,393]

While the relevance of accounting for London dispersion interactions in static computations for excited molecules belonging to categories 2 and 3 has been at least partially recognized, many similar questions remain unanswered for category 1 systems. For instance, how do van der Waals interactions between non-excited large substituents impact excited state processes? Wegner *et al.*[394] offered a first response by demonstrating that the *cis*- to *trans*- thermostability of azobenzene switches is dramatically improved by increasing substituent bulk. While focusing on thermal isomerization process (as opposed to photoisomerization), these results further emphasize the necessity of reevaluating our general perception of steric hindrance, which may be diminished by intramolecular attractive London dispersion forces.[363] This computational work addresses this exact question and explores the role of vdW interactions in the photochemical processes of *cis*-stilbene (**A**) and it substituted 3,3',5,5'-tetra-*tert*-butyl-stilbene (**B**) analogue (Figure 4.2).

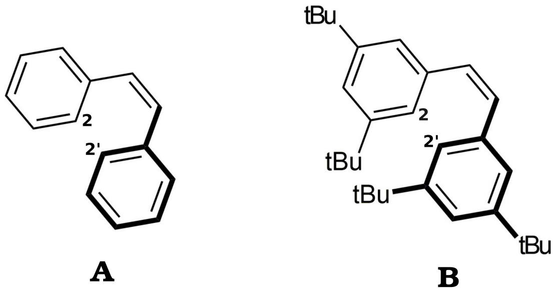


Figure 4.2 – Structure of the compounds studied: **A** cis-stilbene. **B** cis-3,3'-5,5'-tetra-*tert*-butyl-stilbene.

To this end, we compare the static and dynamic excited state energy profiles based on dispersion corrected functionals [PBE0-D3(BJ), PBE0-dDsC, B3LYP-D3(BJ)], their uncorrected variants, as well as a wavefunction-based method (CC2). Our results clearly demonstrate that the effect of dispersion interactions in the excited state process is not negligible.

Stilbene and its derivatives typically undergo reversible photoisomerization processes and 6π-electrocyclization upon exposure to UV light.[395–397] The first singlet excited state corresponds to a fairly pure HOMO-LUMO transition exclusively localized on the stilbene core. In line with the crowded substituted azobenzene derivatives,[394] vdW interactions between the *tert*-butyl substituents in the meta-positions of stilbene strongly influence the potential energy surface of both the S_0 and S_1 states. Specifically, the ground state geometries of **B** optimized at the uncorrected PBE0 level deviates dramatically from those determined at the CC2 level. In contrast, PBE0-D3(BJ), PBE0-dDsC, VV10, and M06-2X geometries agree closely with the wavefunction based method (see structures and RMSD in the supporting information of the original paper). Specifically, the long-range correlation effects between the *tert*-butyl substituents cause the ground state minimum of **B** to be more compact (*i.e.*, a reduced central dihedral angle) than stilbene (**A**), which essentially converge to the same geometry regardless of the optimization level. Relevantly, only the PBE0 (uncorrected) geometries for **B** possess a dihedral angle similar to that of stilbene (**A**). The close similarity of the optimized structures obtained with the four conceptually different methods [PBE0-D3(BJ), VV10, M06-2X and CC2] emphasizes the relevance of London dispersion interactions in shaping the ground state minimum geometry of **B** and, as a result, its Franck-Condon region.

As previously observed,[398,399] the stationary point on the S_1 surface of *cis*-stilbene (**A**) is structurally close to the photocyclization product, 4a,4b-dihydrophenantrene (DHP). Therefore, the distance between the 2 and 2' carbon atoms of the stilbene core is a key variable for characterizing the structural similarity of the relaxed excited state structures of **A** and **B** (supporting information of the original paper). Akin to the ground state, the S_1 minimum of both **A** and **B** computed with TDA-PBE0-D3(BJ) agrees closely with the CC2 optimized structures. In contrast, strong deviations are observed between the TDA-PBE0 S_1 minimum of **B** and the CC2 reference (RMSD ten times larger than with TDA-PBE0-D3(BJ)). In other words, the evolution of **B** away from its Frank-Condon region (FC) toward its S_1 minimum is dictated not only by the driving force of the electronic transition, but also by the existence of attractive dispersion interactions between the large unexcited substituents. This phenomenon is best illustrated by examining the static excited state profiles between the fully optimized S_0 and S_1 minimum (Figure 4.3).

At each point of the profiles, the horizontal axis in Figure 4.3 matches the root mean square deviation (RMSD) between the structures at the TDA-PBE0-D3(BJ) and TDA-PBE0 levels. The choice of coordinates in Figure 4.3 is discussed in more details in Section 5 of the supporting information of the original paper. Overall, the RMSDs allow for the mapping of the energy profiles of both levels on the same horizontal axis. Additionally, the RMSDs encode, directly on the axis, the information relative to the magnitude of the structural differences upon the introduction of the dispersion correction. The two TDA-PBE0 and TDA-PBE0-D3(BJ)) excited state profiles of **A** show no significant deviation in the adiabatic excitation energies. The same trend holds for their geometry, as the RMSD of the mass weighted coordinates falls into a narrow range between 0.01-0.08 Å(x-axis, Figure 4.3, left). In contrast, the excitation energies of **B** using the two levels deviate gradually (up to 0.7 eV, y-axis, Figure 4.3, right) when approaching the minimum structure close to the cyclization product (DHP-like minimum). In particular, the excitation energy of **B** computed with the dispersion corrected scheme, which is lower than for the bare functional, can be explained by the simultaneous occurrence of two physical phenomena: first, upon electronic excitation, the two phenyl rings in both **A** and **B** undergo a conrotatory motion, which brings the 2 and 2' carbon atoms closer. The resulting distortion of the phenyl rings destabilizes the ground state, which, eventually, becomes nearly

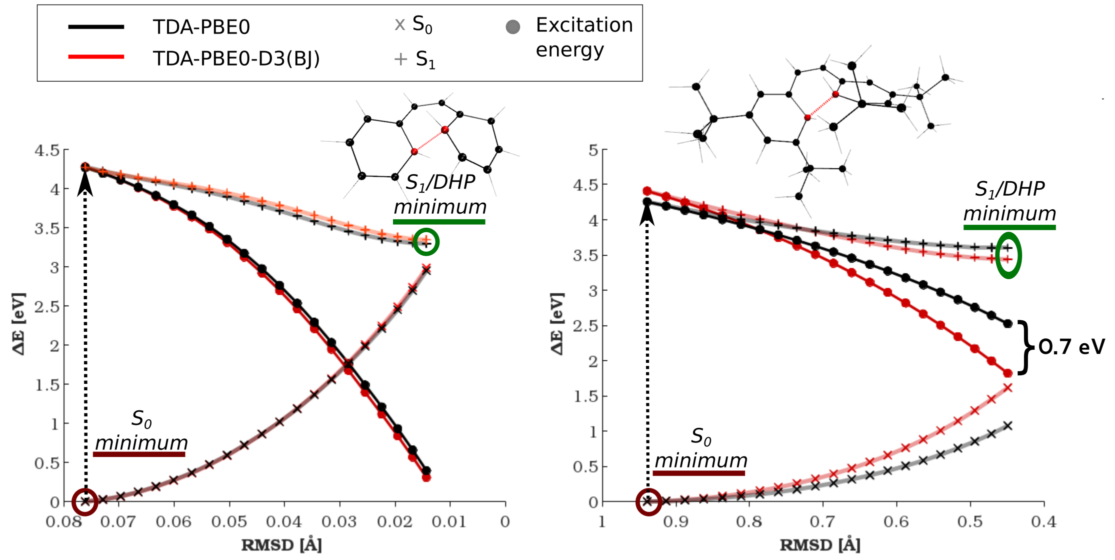


Figure 4.3 – Energy profiles (in eV) of molecule **A** (left) and **B** (right). Faded lines: S_0 and S_1 energies relative to the S_0 minimum. Solid lines: excitation energies. For the red lines, the D3(BJ) dispersion correction was added to the (TDA-)PBE0/def2-SVP level computations.

degenerate with the S_1 state in the region where the photocyclization product is formed. In the case of **B**, however, the attractive interactions between the large *tert*-butyl substituents enhance the rotational motion of the phenyl rings resulting in a more compact structure in which the ground state is even more destabilized. Compared to stilbene (**A**), the presence of the *tert*-butyl substituents depresses the potential energy surface of S_1, resulting in a minimum that is further away from the crossing with the ground state (Figure 4.4, green arrows) with a higher barrier. From a methodological perspective, failure to account for van der Waals interactions results in an additional shift of the S_1 minimum even further away from the crossing region (toward longer C-C distances) and a significant rise in energy of the crossing point with the ground state (purple arrows). Chemically, the presence of *tert*-butyl substituents in **B** should not preclude the accessibility of the photocyclization pathway as followed by stilbene (**A**). In fact, the accessibility of this pathway can only be hindered owing to the limitations of the level of theory chosen for its description (*vide infra*).

Despite being insightful, the static pictures seen in Figures 4.3 and 4.4 misses several key aspects of the main photoreactions that occur within **A** and **B**. If these molecules acquire enough kinetic energy, the potential energy barriers on the S_1 surface of **B** could be overcome and reach the crossing point leading to the photocyclization product (intersection of blue and black lines, Figure 4.4). This should indeed be the case, since it is experimentally known[400] that meta-substituents do not preclude access to the photocyclization pathway of bare stilbene. Excited state adiabatic dynamics simulations are a cost-effective tool that complements the static picture by including effects of kinetic energy and molecular vibrations. We initiated 100 trajectories on the S_1 surface for both **A** and **B**, sampling the ground state configurations from

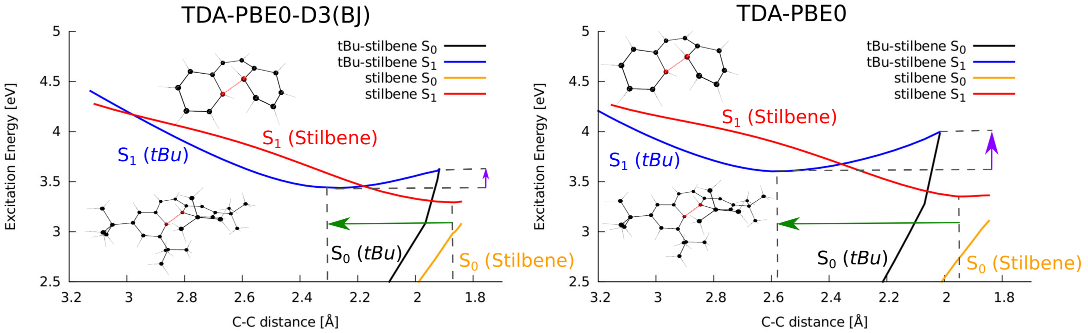


Figure 4.4 – Left panel, TDA-PBE0-D3(BJ). Right panel, TDA-PBE0. Comparison between the S_1 energy profiles of molecules **A** (red and orange lines) and **B** (black and blue lines). Purple arrows: barrier to crossing point with the ground state. Green arrows: shift of the S_1 minimum of **B** with respect to **A**. The energies are relative to the respective S_0 minima.

an independent trajectory run in the canonical ensemble (NVT) at 300 K. While the dispersion correction does not qualitatively alter the evolution of stilbene (**A**) in its S_1 state, different conclusions are drawn for the *tert*-butyl substituted stilbene (**B**). In 18% of the trajectories, a discrepancy in the outcome of the simulation is observed between the dispersion-corrected and uncorrected functionals. As shown in Figure 4.5, the disagreement mainly originates from an increased accessibility to the photocyclization pathway, which arises from the improved description of weak interactions. Provided the inclusion of the D3 dispersion correction, the number of trajectories ending in the photocyclization of molecule **B** increases decisively (Figure 4.5: black points in the blue circle).

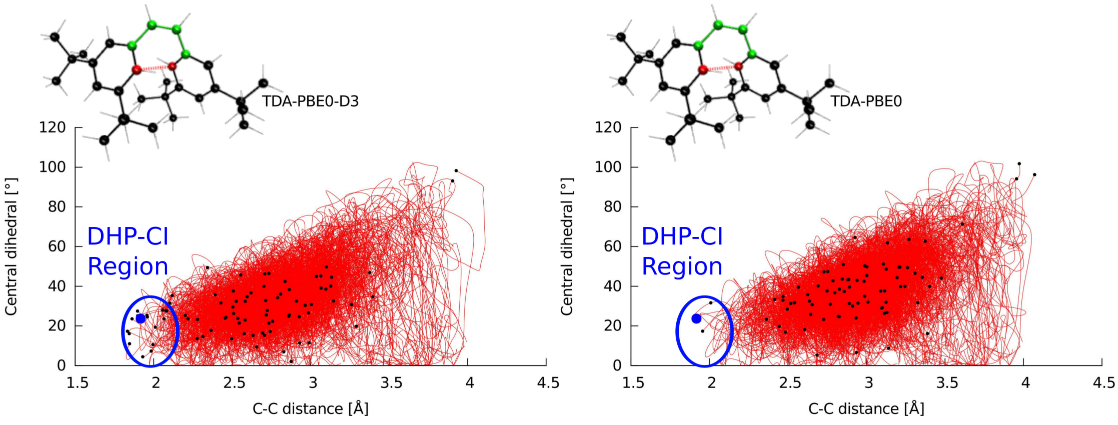


Figure 4.5 – Dynamical evolution of the central dihedral angle (green atoms) and 2-2' carbon-carbon distance (red atoms) on the S_1 surface of the *tert*-butyl substituted stilbene (**B**). Left panel: TDA-PBE0-D3. Right panel: TDA-PBE0. The DHP crossing point region is indicated in blue. The molecular configuration upon the termination of a trajectory is indicated by black dots.

In line with the previously discussed static picture, this result supports the conclusion that a full treatment of van der Waals interactions is critical in capturing the proper evolution of **B** in its first electronic excited state. In this case, London dispersion overpowers repulsive interactions that result from crowding of the *tert*-butyl groups, which allows the system to reach the photocyclization crossing region from the S_1 minimum.

In conclusion, we demonstrated that van der Waals interactions beyond the common ground-state chemical situation cannot be neglected *a priori*. Using molecules stilbene (**A**) and 3,3',5,5'-tetra-*tert*-butyl-stilbene (**B**) as prototypical examples, we found that while the rearrangement of the electronic density upon excitation remains the principle driving force of the excited state processes, London dispersion shapes the potential energy surface and the structural evolution of the photoexcited molecules. This is a crucially important point for excited state molecular dynamic simulations, which may produce erroneous results if the underlying electronic structure method is incapable of accurately treating effects arising from London dispersion forces. From a chemical perspective, comparisons with bare stilbene (**A**) reveal that the substitution with *tert*-butyl groups in **B** results in a stable DHP-like minimum well separated in structure from the crossing region with the ground state. Failure to fully account for van der Waals interactions on the methodological level increases the extent of this structural separation and simultaneously causes a rise in the crossing region energy. The resulting hindrance of the photocyclization pathway for **B** happens solely as a consequence of the failings of the level of theory, not as a result of the chemical substitution with bulky *tert*-butyl groups.

4.1 Computations Methods

The excited states of stilbene (**A**) and *tert*-butyl substituted stilbene (**B**) were computed using linear response TD-DFT within the Tamm-Dancoff approximation (TDA).[401] Single reference methods such as LR-TDDFT cannot provide the correct two-dimensional branching of the conical intersections with the ground state, so that their validity near these regions may be questioned. However, it has been shown that Tamm-Dancoff approximation improves significantly the description of molecules away from their equilibrium geometry and specifically allows to access reliable geometries and energetics in the vicinity of crossing points with the ground state at the TDDFT level (*e.g.* Ref. [402]). Unless explicitly stated otherwise, both single point computations and geometry optimizations were performed with the Turbomole 7.1 program package.[333] The ground and excited-state geometries of molecules **A** and **B** were computed using three global hybrid functionals (PBE0[403,404], B3LYP[82,303,304], BHHLYP[81]) and a range-separate functional (CAM-B3LYP[405]) combined with the def2-SVP[325] basis set. The basis set dependence of the excitation energies was analyzed for both molecule **A** and **B** and resulted in small deviation in the order of 1.75-2.7% of the total excitation energy. Each geometry optimization was performed at first with the bare functional and then including Grimme's D3(BJ) dispersion correction.[160] The vertical excitations and geometry optimizations using the PBE0 functionals were also combined with the dDsC dispersion correction[169–171] as implemented in a development version of QCHEM.[329] CC2 level computations were performed using the resolution of identity and the frozen core approximation. In contrast to LR-TDDFT, CC2 accounts for the effects of non-local correlation,[406] such as van der Waals interactions. Additionally, detailed benchmarking of the method demonstrated its reliability not only near equilibrium geometries, but also in the proximity of conical intersections.[407] Molecular structures between the fully optimized S_0 and S_1 minima were obtained by linear interpolation of the internal coordinates, while the reaction coordinate was scaled using the root mean square deviations (RMSD) of the dispersion corrected structures, relative to those of the uncorrected functional. The adiabatic dynamics simulations were performed using the GPU-accelerated software TeraChem (version 1.9).[408–410] Initial conditions for excited-state dynamics were sampled from a 40 ps long ground state dynamics trajectory within the NVT ensemble (Langevin thermostat 300K). Both ground state and excited state dynamics were computed using the PBE0 functional combined with the 6-31G basis set (for a discussion on the basis set see Supporting Information of the original paper). For each molecule **A** and **B**, two batches of 100 trajectories were initiated on the S_1 surface; the first batch using PBE0 combined with the Grimme's D3 dispersion correction, while the second using the uncorrected functional. All excited-state trajectories were evolved within the NVE ensemble with a time step of 0.5 fs for a maximal time of 1 ps for molecule **A** and 1.5 ps for molecule **B**. If the energy difference between the S_1 and S_0 surfaces dropped below 0.5 eV, the trajectories were terminated, assuming a crossing with ground state surface. Molecular structures were visualized with VMD.[411]

5 Transferable Machine-Learning Model of the Electron Density

This chapter is based on the following publication:

A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, M. Ceriotti, Transferable Machine-Learning Model of the Electron Density, *ACS Cent. Sci.*, **2019**, *5*, 57-64.

with a modified content that places more emphasis on the underlying quantum chemical challenge, *i.e.* the construction of an atom-centered electron density representation compatible with transferable machine-learning.

5.1 Introduction

One of the most compelling consequences of the first Hohenberg-Kohn theorem[221] is that the molecular charge density $[\rho(\boldsymbol{r})]$ contains exactly the same information about any molecular property as the electronic wavefunction. This fundamental equivalence, combined with its simple and unique dependence on real-space variables, have contributed to make the electron density one the most important chemical properties of atoms and molecules. Being a quantum mechanical observable, $\rho(\boldsymbol{r})$ can be obtained experimentally from high-resolution electron diffraction[412,413] and transmission electron microscopy,[414] or alternatively from solving the electronic structure problem through *ab-initio* computations. However, both approaches may become rapidly demanding when the density has to be evaluated for thousands of different molecules/conformations or very large chemical systems.

Machine-learning models are currently emerging as an effective technique to tackle this kind of large scale problems, allowing to bypass the computational cost of *ab-initio* computations. In particular, kernel-based methods are thriving,[41,43] with reported applications ranging from energies,[25] to forces,[37,44,45] spectra,[415], polarizabilities,[47,48] and density functionals.[57] The mathematical complexity of the quantities targeted with statistical learning is rapidly evolving, following the introduction of ever more sophisticated molecular representations and of kernels able to capture fundamental symmetry conservation laws. Yet, the regression of a scalar field such as the electron density remains a non-trivial task.

Predicting the electron density only given a set of nuclear positions represents a challenge for both the regression model and for quantum chemistry. From the machine-learning perspective, it is essential to construct a model able to predict the amplitude of the field simultaneously at every point in space and to capture all the rotational symmetries of the density. From the quantum chemical perspective, it is necessary to provide a representation of the electron density compatible with the learning framework. The most direct approach consists in representing both the molecular structure and the density on a real-space grid. In this way, it is possible to learn and predict $\rho(\boldsymbol{r})$ in each point of the molecular space, at the cost, however, of a particularly intense computational effort.[54,55,416] A different solution has been proposed by Brockherde *et al.*, which have recently built a machine-learning model of the electron density using a smoothed representation of the external potential as fingerprint.[32] In this framework, $\rho(\boldsymbol{r})$ is decomposed onto an orthogonal basis and the molecular information is encoded with a global representation, which allows the construction of many independent kernel-ridge regression models. These choices make the model extremely accurate for small and rather rigid molecules, but also impose substantial constraints to its transferability to large and flexible chemical systems.

In this chapter, we introduce a machine-learning model of the valence electron density that overcomes these limitations. The model is scalable and transferable, *i.e.* it is able to accurately predict $\rho(\boldsymbol{r})$ for large and flexible systems while being trained only on small molecules. The backbone of the model is the choice of an atom-centered, non-orthogonal basis set to represent the density field, in a spirit similar to orbital localization techniques[417–421] and to the multipole analysis of X-ray diffraction.[422–426] The regression of these local density components is performed within a recently introduced symmetry-adapted Gaussian process regression framework, which allows taking advantage of the fundamental symmetries of the decomposed electron density.

The regression model is tested on an ensemble of different conformations of saturated and unsaturated hydrocarbons of increasing size. The accuracy of the learning exercise is analyzed on the smallest molecules including ethene (C_2H_4), ethane (C_2H_6), butadiene (C_4H_6) and butane (C_4H_{10}). As a final result, demonstrating the transferability and the scalability of the model, the valence electron density of octatetraene (C_8H_{10}) and octane (C_8H_{18}) is predicted on the basis of the information gathered from butadiene and butane.

5.2 Expansion of the electron density into an atom-centered, non-orthogonal basis set

The decomposition of molecular properties into additive local contributions is a well-established practice in quantum chemistry, as demonstrated by the abundance of linear-scaling, embedding and fragment decomposition electronic structure methods.[420,427–434] Local and additive partitioning is theoretically justified by the concept of "nearsightedness"[253,254] of all local electronic properties, among which the density is a prototypical example. As already largely shown in the current literature,[25,435–439] additivity and locality are two fundamental qualities to achieve an efficient transferability of machine-learning models.

While several, conceptually different decomposition schemes exist for the electron density,[440,441] none of them can be defined uniquely.[442] Instead of choosing *a priori* a particular scheme, which would hinder the generality of the model, we introduce locality by expanding the density as a sum of atom-centered, non-orthogonal basis functions

$$\rho(\boldsymbol{r}) = \sum_i \rho_i(\boldsymbol{r}) = \sum_{ik} c^i_k \phi^i_k(\boldsymbol{r}) = \sum_{ik} c^i_k \phi_k(\boldsymbol{r} - \boldsymbol{r}_i) = \sum_{inlm} c^i_{nlm} R_n(\boldsymbol{r} - \boldsymbol{r}_i) Y^m_l(\hat{\boldsymbol{r}}_{\boldsymbol{i}}), \tag{5.1}$$

where R_n are Gaussian functions and Y^m_l are spherical harmonics. In this way, it is possible to build a regression model for $\rho(\boldsymbol{r})$, based on the expansion coefficients c^i_{nlm} and the position of the nuclei.

Working with a non-orthogonal basis ensures the efficient transferability of the model, but has the main disadvantages that the prediction of the density expansion coefficients $c^i_{nlm}(\boldsymbol{x})$ cannot be performed independently one from the others as all the density components are coupled by the overlap matrix. This becomes evident when expressing the projections (w^i_k) of the density on the basis functions

$$w^i_k = \langle \rho | \phi^i_k \rangle = \int \mathrm{d}\boldsymbol{r}\ \rho(\boldsymbol{r})\ \phi_k(\boldsymbol{r} - \boldsymbol{r}_i). \tag{5.2}$$

These projections are not simply the expansion coefficients c^i_k, but they are coupled to them by $\mathbf{Sc} = \mathbf{w}$, where $S^{ii'}_{kk'} = \langle \phi^i_k | \phi^i_{k'} \rangle$ is the overlap matrix element between two basis functions. In addition, the evaluation of the integral in Equation 5.2 has to be performed with the highest numerical accuracy, in order to avoid the introduction of spurious noises in the machine-learning model. For instance, integration of the angular part of Equation 5.2 on a regular cubic grid would lead to noisy predictions, resulting in unphysical changes of the electron density upon rotation of a molecule. To avoid this problem, we evaluated the angular part of Equation 5.2 on a particularly dense Lebedev grid, constructed using 2030 angular points.[196,226] The radial integration is less problematic and can be directly performed on an equispaced radial mesh of 200 points spanning a cutoff distance of $r_{\text{cut}} = 6$ Å.

A second important issue in computing the expansion coefficients in Equation 5.1 is the potential ill-conditioning of the overlap matrix. In quantum chemistry, the commonly used strategy to address this problem is the reduction of the size of the radial basis set by contraction of the primitives. Therefore, the final basis set used in the density decomposition has been obtained by the contraction of an initial set of 12 primitives. The contraction coefficients have been optimized by simultaneously minimizing the root mean square density error and the condition-number of the overlap matrix.[443]

5.3 Symmetry-Adapted Gaussian Process Regression for the Electron Density

Encoding the correct rotational symmetries of the electron density in real space is far from being trivial and requires the development of symmetry-adapted regression strategies. This problem has been long known and analyzed in the context of the determination of electron densities by experimental X-ray diffraction.[424–426,444,445] In this field, one of the most widely used approaches describes the valence electron density as a multipolar expansion[446–449] similar in spirit to Equation 5.1. In practice, however, the prediction of the electron density with traditional multipolar models relies on existing pseudoatom libraries, such as ELMAM,[423,450–452] ELMAM2,[453,454] UBDB,[422,455], Invarioms[456] and SBFA[457]. The most straightforward solution in the context of machine-learning would be to choose a fixed frame of reference and align all the tensorial components of the targeted property by rotation and translation.[46,258] Regrettably, this strategy is only applicable for rather small and rigid molecules for which such symmetry operations are well-defined. In contrast, our machine-learning model aims at describing the electron density of arbitrarily complex and flexible molecules.

Gaussian process regression (GPR) can be reformulated to account for the symmetries of the SO(3) group, which has been already demonstrated for the regression of vectors[458] and tensors of any order.[47] This symmetry-adapted-Gaussian-process-regression (SA-GPR) framework can be readily applied to a density decomposition expressed as in the previous section, where the prediction of the density components becomes,

$$c^i_{nlm}(\boldsymbol{x}) = \sum_{j\in M}\sum_{|m'|\leq l} k^l_{mm'}(\mathcal{X}_i,\mathcal{X}_j)\; x^j_{nlm'}\delta_{\alpha_i\alpha_j} \tag{5.3}$$

where, $\boldsymbol{k}^l(\mathcal{X}_i,\mathcal{X}_j)$ is a kernel matrix of dimension $(2l+1)\times(2l+1)$ that expresses both the structural similarity and the geometric relationship between an atom-centered environment $\mathcal{X}_i$ of the target molecule and a set M of reference environments $\mathcal{X}_j$. The regression weights $x^j_{nlm'}$ are determined from a set of N training configurations and their associated electron densities.

In particular, the choice of spherical harmonics as the angular basis in Equation 5.1 allows regrouping the coefficients according to their angular momentum l in a set of spherical multipoles $\boldsymbol{c}^i_{nl}$ of dimension $2l+1$. In this way, the density can be decomposed in a sum of spherical tensor components with rank l, beginning with the spherical elements ($l=0$) and increasingly adding anisotropy ($l>0$). Restraining the model to $l=0$ would result in a predicted density similar in spirit to the widely used promolecular approach,[459] or superposition of spherical atom densities (SAD).[193]

Upon the accurate evaluation of Equation 5.2, the coefficients could be, in principle, simply determined by multiplication with the inverted overlap matrix. However, this direct procedure led to rather poor regression accuracy and unstable predictions. The reason behind this poor performance is rooted in the fact that the overlap matrix **S** is often ill-conditioned, which dramatically amplifies any arbitrarily small numerical error done in the evaluation of $\boldsymbol{w}$. To overcome this problem and improve the accuracy of the model, the basis set decomposition and the construction of the machine-learning model have been combined in a single step. In other words, we require the regression model not only to predict the electron density, but also to find, out of the many nearly equivalent decompositions of ρ, the one which best fits the target density associated with a given structure.

The problem can be cast into a single least-square optimization of a loss function that measures the discrepancy between the reference and the model densities,

$$\mathscr{L}(\boldsymbol{x}) = \sum_{\mathscr{A}\in N} \int d\boldsymbol{r} \left| \rho_{\mathscr{A}}(\boldsymbol{r}) - \sum_{i\in\mathscr{A}} \sum_{k} c_k^i(\boldsymbol{x})\, \phi_k(\boldsymbol{r}-\boldsymbol{r}_i) \right|^2 + \eta\, |\boldsymbol{x}|^2 . \tag{5.4}$$

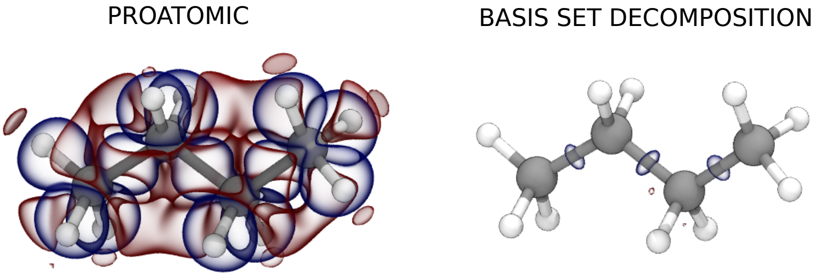
where N runs over the training set and i runs over the environments of a given training structure. The second term in the loss is a regularization, which avoids overfitting. In this context, η represents an adjustable parameter that is related to the intrinsic noise of the training dataset.

The coefficients $\boldsymbol{c}$ depend parametrically on the regression weights $\boldsymbol{x}$ via Eq. (5.3); by differentiating the loss with respect to x_{nlm}^j one obtains a set of linear equations that make it possible to evaluate the weights in practice. In compact notation, the solution of this problem reads

$$\boldsymbol{x} = \left(\boldsymbol{K}^T \boldsymbol{S} \boldsymbol{K} + \eta \mathbf{1}\right)^{-1} \boldsymbol{K}^T \boldsymbol{w} \tag{5.5}$$

where $\boldsymbol{x}$ and $\boldsymbol{w}$ are vectors containing the regression weights and the density projections on the basis functions, while $\boldsymbol{K}$ and $\boldsymbol{S}$ are sparse matrix representations containing the symmetry-adapted tensorial kernels and the spatial overlaps between the basis functions. The similarity measures in the regression formula (Equation 5.5) are the λ-SOAP kernels (see Chapter 2).[47] These kernels are a generalization of the scalar ($\lambda = 0$) smooth overlap of atomic positions framework[59] that has been used successfully in the construction of interatomic potentials[436,460] and in the prediction of molecular properties.[461,462]

It should, however, be stressed that the final regression problem is highly non-trivial. The kernels that involve environments within the same training configuration are coupled by the overlap matrix, so that all the regression weights $\boldsymbol{x}$ for different elements, radial and angular momentum values must be determined simultaneously. An efficient implementation of an ML model based on Equation (5.5) requires the optimization of a basis set for the expansion, the evaluation of $\rho(\boldsymbol{r})$ on dense atom-centered grids, the sparsification of the descriptors that are used to evaluate the kernels, and the determination of a diverse, minimal set of reference environments X_j.



$\langle\varepsilon_\rho\rangle$(%)	C_2H_4	C_2H_6	C_4H_6	C_4H_{10}
Proatomic	18.06	19.23	16.79	18.13
Basis Set	1.04	1.14	0.98	1.19

Figure 5.1 – Density errors at different level of representation: (*left*) superposition of isolated atomic densities, (*right*) optimized basis set. Red and blue isosurfaces refer to an error of ±0.005 Bohr^{-3} respectively. The density errors for the structure depicted are reported in the two panels, while the table reports the mean errors over the whole training set for the C_2 and C_2 molecules.

5.4 Results and Discussion

5.4.1 Charge decomposition analysis

The main drawback of expanding the density into local contributions is the introduction of a decomposition error along with the intrinsic prediction error related to the statistical nature of machine-learning. With a basis set of 4 contracted radial functions and angular momentum components up to $l = 3$, the typical error in the density decomposition can be brought down to about 1%. To develop a more intuitive understanding of the relative magnitude of the decomposition error, we compare it to the error that can be expected using promolecular densities, which ranges from 16 to 20% for the different molecules in the dataset (Figure5.1).

The form of the Ansatz for the density decomposition (Equation 5.1) allows the individual analysis of each angular momentum channel l. This is shown in Figure,5.2 where the isotropic $l = 0$ functions largely determine the general shape of the density. Nevertheless, higher angular momenta are crucial to describe fundamental anisotropies of the electron density: the $l = 1$ functions capture the gradient of electronegativity in the C–H bonds, the $l = 2$ functions describe the σ and π system of the C–C bonds along the main chain, while $l = 3$ absorb all the other non-trivial anisotropies. The bottom panel of the Figure shows the contribution to the valence electron density of angular momentum l and atomic type α, *i.e.* $\sigma(l,\alpha) = \sqrt{\sum_n \langle |\mathbf{c}^i_{ln} - \langle \mathbf{c}^i_{ln}\rangle|^2\rangle_{\alpha_i=\alpha}}$, with the average $\langle\cdot\rangle$ involving all the atoms of the same type included in the dataset. After baselining the valence electron density by subtraction of the mean atomic density of pure $l = 0$ character, the $l = 1$ components largely dominate the charge density variability associated with hydrogen atoms. Higher angular momenta also carry a

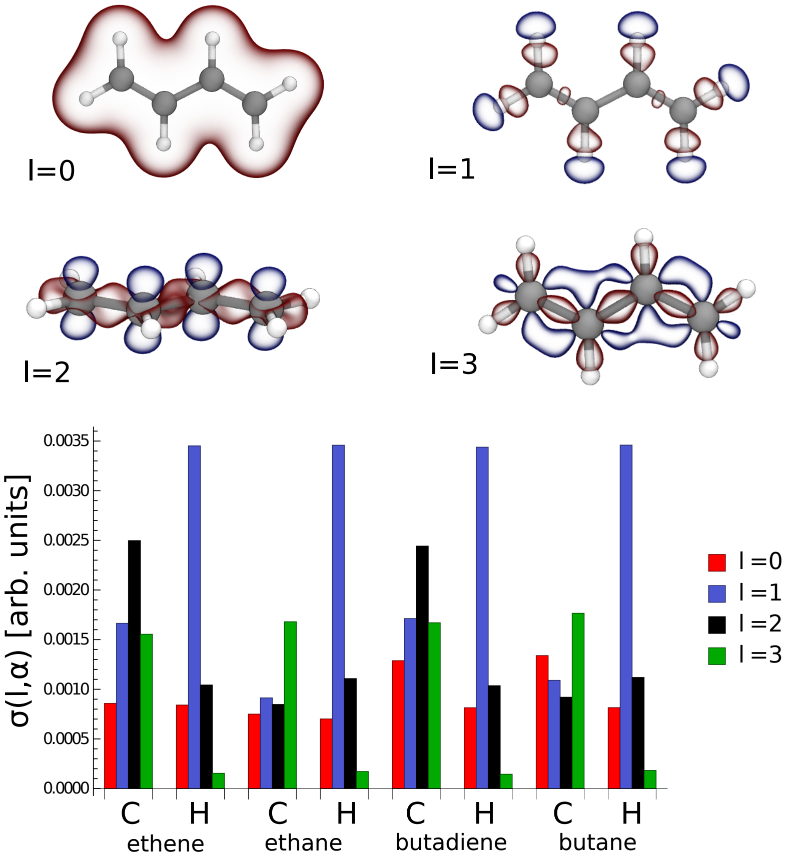


Figure 5.2 – (*top*) representation of the angular momentum decomposition of the electron density. Red and blue isosurfaces refer to ±0.01 Bohr^{-3} respectively. (*bottom*) angular momentum spectrum of the valence electron density of C_2 and C_4 datasets. The isotropic contributions $l = 0$ express the collective variations with respect to the dataset's mean value, while the mean is statistically zero for $l > 0$.

substantial contribution to the description of the carbon atom density in alkenes ($l = 2$) and alkanes ($l = 3$), in agreement with the existing literature.[423]

5.4.2 Density learning with SA-GPR

The performance of the machine-learning model in terms of the prediction accuracy of the electron density as a function of the number of training molecules is shown in Figure 5.3. The structural flexibility of the molecules largely controls the accuracy of the final machine-learning model. The error in the smallest systems, such as ethene and ethane, decreases rapidly, showing that these molecules could be perfectly learned in a less-sophisticated framework based on the alignment to a fixed frame of reference. On the other hand, butadiene is more challenging both because of its greater conformational variability (*e.g.*, *cis* and *trans* conformers, as well as distorted configurations approaching the isomerization transition-state) and because its π-density is more sensitive to small molecular deformations. The difficulty of

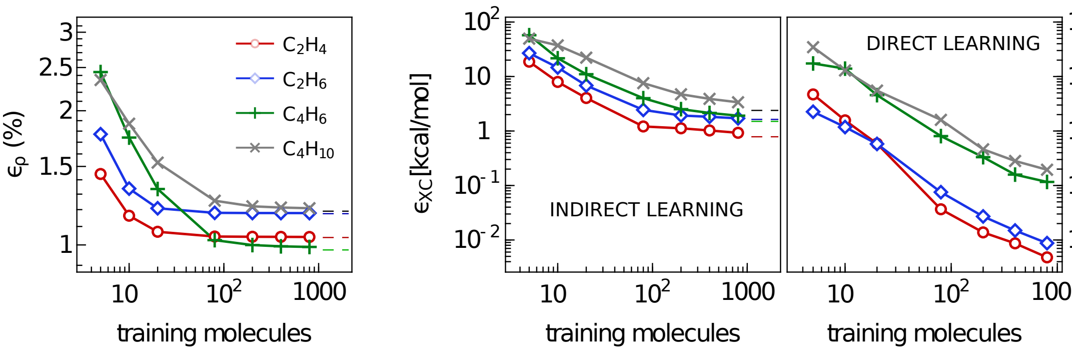


Figure 5.3 – Learning curves for C_2 and C_4 molecules. (*left*) % mean absolute error of the predicted SA-GPR densities as a function of the number of training molecules. The error normalization is provided by the total number of valence electrons. (*right*) root mean square errors of the exchange-correlation energies indirectly predicted from the SA-GPR densities and directly predicted via a scalar SOAP kernel, as a function of the number of training molecules. Dashed lines refer to the error carried by the basis set representation.

the learning exercise is increased for butane due to its flexibility and the broad spectrum of intramolecular non-covalent interactions spanned by the many different conformers contained in the dataset. For these reasons, this kind of system is expected to benefit most from an ML scheme that can adapt its kernel similarity measure to different orientations of molecular fragments.

The learning curves are obtained by varying the number of training molecules up to 800 randomly selected configurations out of the total of 1000. The remaining 200 molecules for each of these random selections are used to estimate the error in the density prediction. The measure of the error is the mean absolute difference between the predicted and quantum mechanical densities, *i.e.*, $\varepsilon_\rho(\%) = 100 \times \langle \int d\boldsymbol{r} \, \left|\rho_{\mathrm{QM}}(\boldsymbol{r}) - \rho_{\mathrm{ML}}(\boldsymbol{r})\right| \rangle / N_{\mathrm{e}}$.

The accuracy of the model in ethene and ethane is limited by the error of the basis set representation (around 1% for all molecules) and decreases rapidly, converging to saturation with only 10 training molecules. Due to their increased complexity and conformational flexibility, butadiene and butane are more challenging, but eventually also reach the basis set limit with 100 training structures. While this level of accuracy was demonstrated to be sufficient for density applications in real-space,[459] using the predicted $\rho(\boldsymbol{r})$ to compute exchange-correlation energies remains a challenge. In particular, evaluating the PBE exchange-correlation functional $E_{XC}[\rho]$ with the SA-GPR predictions results in root mean square errors of 0.9 and 1.7 kcal/mol for ethene and ethane, 1.9 kcal/mol for butadiene and 3.5 kcal/mol for butane with the full training set. Again, the performance of the model is limited by the underlying basis set decomposition error, suggesting that the predictions could potentially reach a higher overall accuracy. Nevertheless, it has to be noted that, as far as the exchange-correlation energy is concerned, adopting a direct, conventional scalar regression should lead to vastly superior performance while requiring a much simpler machine-learning scheme (Figure 5.3 rightmost panel).

5.4.3 Size-extensive extrapolation

While the choice of a local basis set decomposition inevitably introduces additional errors in the regression, it is the backbone of the scalability and the transferability of the model. In practice, the locality of the basis implies that the regression weights x^j_{nlm} can be learned from any compound and used as building blocks to predict the density of much more complex molecules. As long as the training set contains enough chemical diversity to capture all the possible local environments of the desired target, the density can be directly obtained by computing the similarity measures ($\boldsymbol{k}^l(\mathscr{X}_i,\mathscr{X}_j)$) between the target molecule ($\mathscr{X}_i$) and the reference training environments ($\mathscr{X}_j$). The computational cost of such prediction is proportional to the number of atom-centered environments, which makes the model strictly linear scaling in the size of the target molecule.

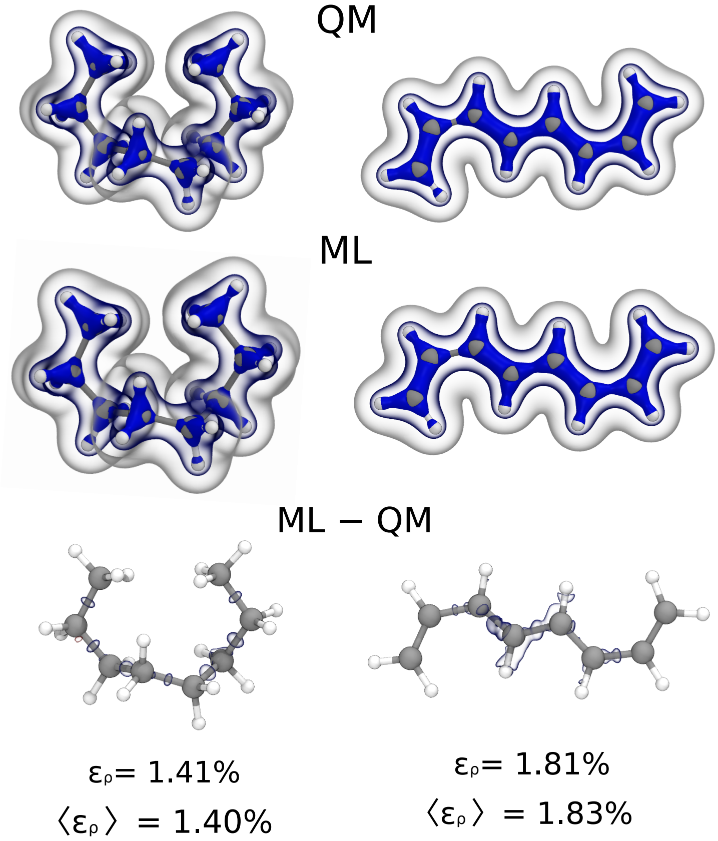


Figure 5.4 – Extrapolation results for the valence electron density of one octane (*left*) and one octatetraene (*right*) conformer. (*top*) DFT/PBE density isosurface at 0.25, 0.1, 0.01 Bohr^{-3}, (*middle*) machine-learning prediction isosurface at 0.25, 0.1, 0.01 Bohr^{-3}, (*bottom*) machine-learning error, red and blue isosurfaces refer to ± 0.005 Bohr^{-3} respectively. Relative mean absolute errors averaged over 100 conformers are also reported for both cases.

To demonstrate the accuracy of the extrapolation, we predict the valence electron density of octatetraene and octane, using only the reference environments and the regression coefficients trained on the butadiene and butane data. While the term extrapolation perfectly captures the increased complexity of the targets with respect to the training molecules, the procedure is, in fact, an interpolation in the space of local environments. For instance, an optimal extrapolation accuracy is obtained using a machine learning cutoff of $r_{\text{cut}} = 3$ Å, instead of $r_{\text{cut}} = 4.5$ Å used for same-molecule predictions. The reduction of the optimal cutoff radius is not only a technical issue, but it shows that beyond 3 Å the environments of octane and octatetraene differ substantially from those in the corresponding C4 compounds. In the ideal case, reducing the representation cutoff radius could be avoided by extending the training set to include molecules of a size comparable to the extrapolation targets.

For both octane and octatetraene, the extrapolation is carried out on a dataset made of the 100 most diverse structures extracted by farthest point sampling from the 300 K replica of a long REMD run. The final accuracy of the model, obtained for the using the full dataset (training and test) of butadiene and butane, reaches a mean absolute percentage error of 1.8% for octatetraene and of 1.4% for octane. As shown in Figure 5.4 for two representative configurations, the extrapolation model reproduces quite accurately the general structure of the valence electron density of both octane and octatetraene. The largest error occurs in the middle regions of octatetraene, for which no analogous examples are contained in the butadiene training dataset.

Besides these prototypical examples, the SOAP representation can be extended to more complex molecules and condensed phases[463], and has been shown to achieve an impressive accuracy predicting the properties of larger molecules while training on very simple compounds[48]. Obtaining similar results for the electron density involves some technical challenges, connected with the presence of correlations between coefficients due to the non-orthogonal basis expansion, which makes the cost of training (but not of predicting) the density scale unfavorably with system size.

5.5 Conclusions

The work described in this chapter demonstrates how accessing some of the most fundamental molecular properties, such as the electron density, requires the development of technically sophisticated machine-learning models. Aiming at transferability across molecules of different size and composition, a scheme has to be local and should account for the fundamental physical symmetries of the problem without any other *a priori* assumption. Our model fulfills these requirements by decomposing the density in optimized atom-centered basis functions, using a symmetry-adapted regression scheme, and designing a loss function that relies only on the total electron density as a physically-meaningful constraint.

The performance of our model was trained and evaluated on a dataset of ground-state valence electron densities of simple saturated and unsaturated hydrocarbons, achieving in all cases an error of the order of 1% on the predicted density. Given its accuracy and its simple dependence only on nuclear positions, the model could find several different chemical applications, such as in the analysis of X-ray[423] and transmission electron microscopy experiments.

Finally, the model could be further improved by, among others, better choice of the decomposition scheme and of the basis set or using computationally cheap semi-empirical methods to provide a baseline for the electron density prediction in a Δ-learning framework. In fact, the work presented in this chapter has to be seen as a first successful attempt to apply transferable machine learning to molecular properties with complex fundamental symmetries. Other examples of properties that would benefit from such an approach are the Hamiltonian and the density matrix, vector fields and density response functions. The wide variety of chemical applications spanned by these properties emphasizes the potential impact of symmetry-adapted Gaussian process regression applied to fundamental quantum chemical properties.

5.6 Computational Details

As a demonstration of our framework, we consider hydrocarbons, using a dataset of 1000 independent structures of ethene, ethane, butadiene, and butane. Atomic configurations are generated by running replica exchange molecular dynamics (REMD) simulations at the density functional tight binding level[464], using a combination of the DFTB+[465] and i-PI[466] simulation software.[467] In order to construct a realistic and challenging test of the ML scheme, we chose the replica at $T = 300$ K and selected a diverse set of 1000 configurations, by a farthest point sampling (FPS) algorithm based on the SOAP metric[461,468]. For each selected configuration we computed the valence electron pseudo density at the DFT/PBE level with SBKJC effective core potentials.

The problem of representing a charge density in terms of a non-orthogonal localized basis set shares many similarities with that of expanding the wavefunction. For this reason, we resort to many of the tricks used in quantum chemistry codes, including the use of Gaussian type orbitals (GTOs) to compute the basis set overlap analytically, and the contraction of 12 regularly spaced radial GTOs down to 4 optimized functions. We find that angular momentum channels up to $l = 3$ functions are needed to obtain a decomposition error around 1% for the density. The coefficients of the contraction are optimized to minimize the mean charge decomposition error and the condition number of the overlap matrix for the four molecules.[443]

A systematic analysis of the interplay between the details of the basis set and the performance of the ML model goes beyond the scope of this work. It is likely however that substantial improvements of this approach could be achieved by further optimization of the basis.

6 Electron density learning of non-covalent systems

This chapter is based on the following publication:

A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, C. Corminboeuf, Electron density learning of non-covalent systems, *Chem. Sci.*, **2019**,*10*, 9424-9432.

6.1 Introduction

Non-covalent interactions (NCIs) govern a multitude of chemical phenomena and are key components for constructing molecular architectures.[469] Their importance fostered an intense research effort to accurately quantify their magnitude and develop an intuitive characterization of their physical nature using quantum chemistry.[470–474] Among the different approaches to characterize non-covalent interactions, one of the simplest and most generally applicable takes as a starting point the electron density $\rho(\boldsymbol{r})$ that encodes, in principle, all the information needed to fully characterize a chemical system[221]. Despite the fact that the universal functional relationship between total energy and $\rho(\boldsymbol{r})$ remains unknown, existing approximations within the framework of Kohn-Sham DFT (KS-DFT)[4] do permit access to all molecular properties within a reasonable degree of accuracy.[7,8,11]

Properties that can be derived exactly from the electron density distribution include molecular and atomic electrostatic moments (*e.g.*, charges, dipole, quadrupoles), electrostatic potentials and electrostatic interaction energies. Knowledge of these quantities is fundamental in diverse chemical applications, including the computation of the IR intensities[475], the identification of binding sites in host-guest compounds,[476–478] and the exact treatment of electrostatics within molecular simulations.[479] Moreover, analyzing the deformation of $\rho(\boldsymbol{r})$ in the presence of an external field provides access to another set of fundamental properties, namely molecular static (hyper)polarizabilities and, thus, to the computation of Raman spectra[480] and non-linear optical properties.[481–484]

The natural representation of the electron density in real space makes it especially suitable for accessing spatial information about structural and electronic molecular properties, including

X-Ray structure refinement[485–490] and representations using scalar fields.[474] Routinely used examples include the quantum theory of atoms in molecules (QTAIM),[491,492] the density overlap region indicator (DORI),[493] and the non-covalent interaction (NCI) index.[459,494]

$\rho(\boldsymbol{r})$ is generally obtained by solving the electronic structure problem through *ab-initio* computations. The main advantage of this approach is that it returns the variationally optimized electronic density for a given Hamiltonian. Yet, *ab-initio* computations can become increasingly burdensome if $\rho(\boldsymbol{r})$ has to be evaluated for thousands of different molecules or for very large chemical systems, such as peptides and proteins. These large scale problems are typically tackled using a more scalable approach that consists of either using linear scaling techniques such as Mezey's molecular electron density LEGO assembler (MEDLA)[495,496] and adjustable density matrix assembler (ADMA)[497–499], as well as approaches based on localized molecular orbitals, such as ELMO.[500–503] Another methodology belonging to this second category involves the use of experimental techniques, such as X-Ray diffraction, to probe the electron density and subsequently reconstructing $\rho(\boldsymbol{r})$ through multipolar models[504–506] and pseudo-atomic libraries, such as ELMAM,[423,450–452] ELMAM2,[453,454] UBDB,[422,455] Invarioms[456] and SBFA.[457] While successful, these two methodologies have intrinsic limits: the first is unable to capture the deformations of the charge density due to intermolecular interactions unless a suitable fragment is generated *ad-hoc*, while the second relies on experimental data and is difficult to extend to thousands of different chemical systems at once. Recently, the development of several machine-learning models targeting the electron density has effectively established a third promising methodology, with the potential to overcome the limitations of the more traditional approaches.

The first machine-learning model of $\rho(\boldsymbol{r})$ was developed on the basis of the Hohenberg-Kohn mapping between the nuclear potential and the electron density.[32,507] Although successful, the choice of the nuclear potential as a representation of the different molecular conformations and the expansion of the electron density in an orthogonal plane-wave basis effectively constrained this landmark model to relatively small and rigid molecules with limited transferability to larger systems. Recently, we proposed an atom-centered, symmetry-adapted Gaussian process regression[47] (SA-GPR) framework explicitly targeting the learning of the electron density.[33] Using an optimized non-orthogonal basis set, pseudo-valence electron densities could be predicted in a linear-scaling and transferable manner, meaning that the model is able to tackle much larger chemical systems than the those used to train the regression model. A third approach, that can also achieve transferability between different systems, uses a direct grid-based representation of the atomic environment to learn and predict the electronic density in each point of the molecular space.[54,55,416] Representing the density field on a large set of grids points rather than on a basis set effectively avoids the introduction of a basis set error, but also dramatically increases the computational effort.

One should also consider that machine learning, being a data-driven approach, requires high-quality, diverse reference data. Fortunately, several specialized benchmark databases that target NCIs have appeared over the past decade. From the original S22[266] to NCIE53[140], S66[271], NBC10/NBC10ext[273–275], and S12L[272,508], the evolution of these datasets has, generally, followed a prescription of increasing the number of entries, principally by including subtler interactions and/or larger systems. In this respect, the databases of Friesner,[278] Head-Gordon,[280] Shaw,[279] and the recent BFDb of Sherrill,[61] constitute a special category because of their exceptional size (reaching thousands of entries) which are now sufficiently large to be compatible with machine-learning applications. Beyond their conceptual differences, each of these benchmark sets aim at improving the capability of electronic structure methods to describe the energetic aspects of non-covalent interactions.

In this chapter, we introduce a dramatic improvement of our previous density-learning approach by making the regression machinery of $\rho(\boldsymbol{r})$ compatible with density-fitting auxiliary basis sets. These specialized basis sets are routinely used in quantum chemistry to approximate two-center one-electron densities. Here, the auxiliary basis sets are used directly to represent the electron densities that enter our machine-learning model, with the additional advantage of avoiding the arbitrary basis set optimization procedures on the machine-learning side. This enhanced framework leverages the transferability of our symmetry-adapted regression method and is capable of learning the all-electron density across a vast spectrum of 2291 chemically diverse dimers formed by sidechain-sidechain interactions extracted from the BioFragment Database (BFDb)[61]. The performance of the method is demonstrated through the reproduction of $\rho(\boldsymbol{r})$ between and within each monomer forming the dimers. The accuracy of the predicted densities is assessed by computing density-based scalar fields and electrostatic potentials, while the errors made with respect to the reference densities are computed by direct integration on three-dimensional grids. As a major breakthrough, the model is used to predict the charge density of a set of 8 polypeptides (~100 atoms) at DFT accuracy in few minutes.

6.2 Methods

Gaussian process regression (GPR) can be extended to encode all the fundamental symmetries of the $O(3)$ group, effectively allowing machine-learning of all the molecular properties that transform as spherical tensors under rotation and inversion operations.[47,509] In the specific case of the electron density, the scheme relies upon the decomposition of the field into additive, atom-centered contributions and the subsequent prediction of the corresponding expansion coefficients.[33] In SA-GPR, each molecule is represented as a collection of atom-centered environments, whose relationships and similarities are measured by symmetry adapted kernels. An in-depth discussion about how a symmetry adapted regression model of the electron density can be constructed is reported in the supplementary materials.

The decomposition of the electron density in continuous atom-centered basis functions is the cornerstone of the scalability and transferability of our SA-GPR model. Beside being generally desirable, these properties are actually crucial to accurately describe the chemical diversity present in the BioFragment Database within a reasonable computational cost. On the other hand, the projection of the density field onto a basis set leads to an additional error on top of that which can be ascribed to machine learning. In practice, all the efforts placed into achieving a negligible machine-learning error are futile if the overall accuracy of the model is dictated by a large basis set decomposition error.

Standard quantum chemical basis sets are generally optimized to closely reproduce the behavior of atomic orbitals[2] and results in unacceptable errors if used to decompose the electronic density (Figure 6.1). In contrast, specialized basis sets used in the density fitting approximation (also known as resolution-of-the-identity (RI) approximation)[204–206,211,214,510,511] are specifically optimized to represent a linear expansion of one-electron charge densities obtained from the product of atomic orbitals. Using the RI-auxiliary basis sets $\{\phi_k^{RI}\}$, the total electron density field can be expressed as:

$$\rho(\boldsymbol{r}) = \sum_{k}^{N_{aux}} \left(\sum_{ab}^{N_{AO}} D_{ab}\, d_k^{ab} \right) \phi_k^{RI}(\boldsymbol{r}) = \sum_{k}^{N_{aux}} c_k \phi_k^{RI}(\boldsymbol{r}) \tag{6.1}$$

where, D_{ab} is the one-electron reduced density matrix and d_k^{ab} are the RI-expansion coefficients. Given a molecular geometry, the value of the basis functions can be readily computed at each point of space, leaving the c_k expansion coefficients as the only ingredient needed by the machine-learning model to fully determine $\rho(\boldsymbol{r})$ (more details in the ESI).

As shown in Figure 6.1, the use of the RI-auxiliary basis sets results in nearly two orders of magnitude increase in the overall accuracy with respect to the corresponding standard basis set. The addition of diffuse functions marginally improves the performance of the decomposition, but leads to instabilities of the overlap matrix (high condition number) and increases dramatically the number of basis functions per atom.

In practice, Weigend's cc-pVQZ/JKFIT[204] basis set (henceforth: cc-pVQZ-RI) offers the best trade-off between accuracy and computational demand and therefore represents the best choice for the density decomposition.

6.2.1 Computational Details

The dataset of molecular dimers has been selected from the side-chain side-chain interaction (SSI) subset of the BioFragment Database (BFDb).[61] The original set is made of 3380 dimers formed by amino-acids side-chain fragments taken from 47 different protein structures. Dimers with more than 25 atoms as well as those containing sulfur atoms were not considered. While the total number of sulfur-containing structures is too small to enable the machine-learning model to accurately capture its rich chemistry, the inclusion of the larger systems does not increase dramatically the chemical diversity of the dataset. The final dataset contains a total of 2291 dimers.

As shown in Figure 6.2, the complete set of 2291 dimers spans a large variety of dominant interaction types, ranging from purely dispersion dominated complexes (in blue) to mixed-

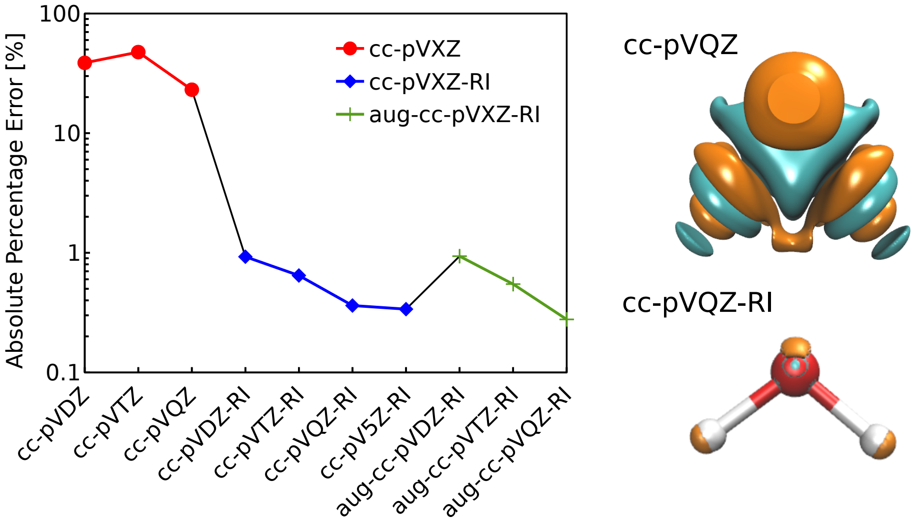


Figure 6.1 – (*left*) Decomposition error of the electron density of a single water molecule: evolution of the absolute percentage error depending on the choice of decomposition basis set. (*right*) Comparison of the density error made with the standard and the RI-auxiliary cc-pVQZ basis set (cyan and orange isosurfaces refer to an error of ±0.005 Bohr^{-3}). Reference density: PBE/cc-pVQZ.

influence (green and yellow) to hydrogen-bonded and charged systems (red). We retain the same classification criteria as in the original database to attribute the nature of the dominant interaction.

For each dimer, the reference full-electron density has been computed at the ωB97X-D/cc-pVQZ level using the resolution of identity approximation for the Coulomb and exchange potential (RI-JK). This implies that RI-auxiliary functions up to $l = 5$ are included for carbon, nitrogen and oxygen atoms while auxiliary functions up to $l = 4$ are used for hydrogen atoms.

6.3 Results and Discussion

The training set for the density-learning model was chosen by randomly picking 2000 dimers out of a total of 2291 possibilities. The remaining 291 were used to test the accuracy of the predictions. Given the tremendous number of possible atomic environments (~40 000) associated with such a chemically diverse database, a subset of M reference environments was selected to reduce the dimensionality of the regression problem (see Supplementary information). To assess the consequences of this dimensionality reduction, the learning exercise was performed on three different sizes $M = \{100, 500, 1000\}$ for the reference atomic environments.

Figure 6.3 summarizes the performance of the machine learning algorithm, expressed in terms of the mean absolute difference between the predicted and the reference densities

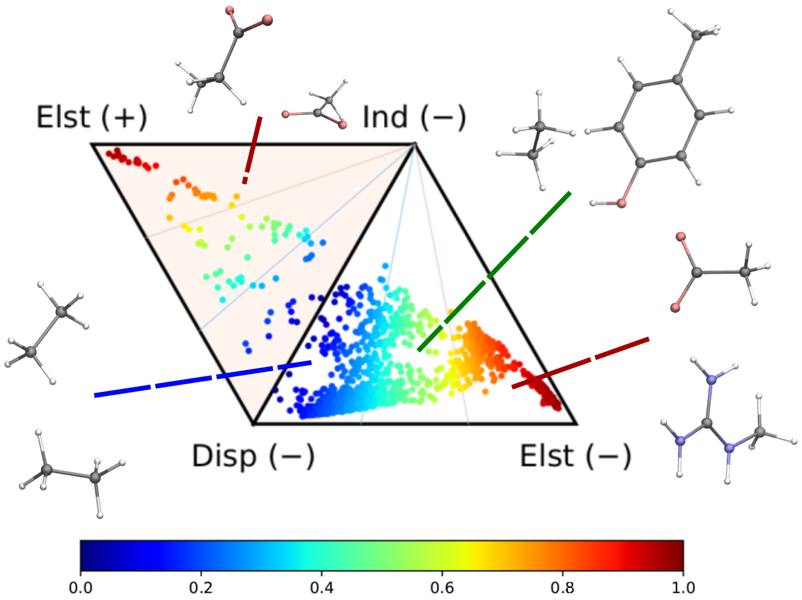


Figure 6.2 – Ternary diagram representation of the attractive components of the dimer interaction energies for the 2291 systems considered in this chapter. The values of the SAPT analysis are taken from Ref. 61.

(QM). Here, only the machine-learning error is shown as the reference densities derive from the RI-expansion of the computed *ab-initio* densities. Since the test set contains molecules of different sizes, the contribution of each dimer has been weighted considering the ratio between its number of electrons and the total number of electrons in the test set.

$$\epsilon_\rho(\%) = 100 \times \frac{1}{N_e}\sum_i N_e^i \frac{\int d\boldsymbol{r}\, \left|\rho^i_{QM}(\boldsymbol{r}) - \rho^i_{ML}(\boldsymbol{r})\right|}{\int d\boldsymbol{r} \rho^i_{QM}(\boldsymbol{r})} \tag{6.2}$$

where the sum is performed over the 291 dimers of the test set, N_e is the total number of electrons, N_e^i is the number of electrons in a dimer, $\rho^i_{QM}(\boldsymbol{r})$ and $\rho^i_{ML}(\boldsymbol{r})$ are, respectively, the *ab-initio* and the predicted density amplitudes at a point. Both integrals of Eq. 6.2 are evaluated in real-space over a cubic grid with step size of 0.1 Bohr in all direction and at least 6 Å between any atom and the cube border.

As shown in the first panel of Figure 6.3, 100 training dimers were sufficient to reach saturation of the density error around 0.5% for M=100. This result already outperforms the level of accuracy reached in our previous chapter, which is remarkable given the large chemical diversity of the dataset and the consideration of all-electron densities. Learning curves obtained with M=500 and M=1000 show steeper slopes, approaching saturation at about 2000 training dimers with errors that were reduced to ~0.2-0.3%. The predicted full-electron den-

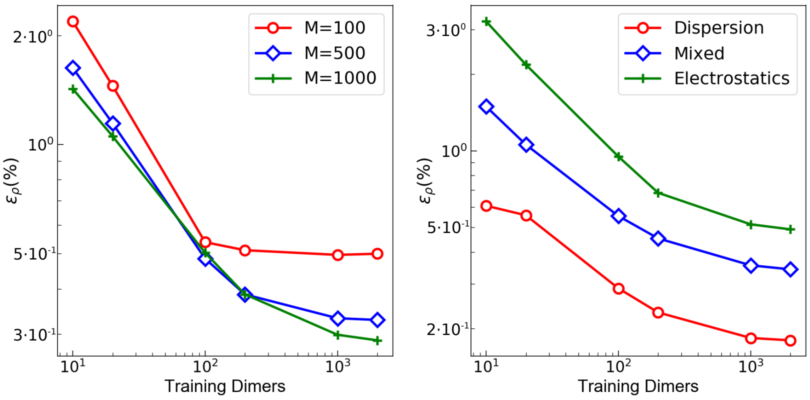


Figure 6.3 – Learning curves with respect to RI-expanded densities (ML error). (*left*) weighted mean absolute percentage error ($\epsilon_\rho(\%)$) of the predicted SA-GPR densities as a function of the number of training dimers. The weights correspond to the number of electrons in each dimer and the normalization is defined by the total number of electrons. Color code reflects the number of reference environments. (*right*) $\epsilon_\rho(\%)$ of the predicted SA-GPR densities (M=1000) divided per dominant contribution to the interaction energy according to Ref. 61.

sities are five times more accurate than the previous predictions of valence-only densities (approximately 1%).[33] A more detailed analysis of the M=1000 learning curve reveals a strong dependence on the nature of the dominant interaction (Figure 6.3). Specifically, stronger non-local character in the interaction yields a larger error. This is especially prevalent for dimers dominated by electrostatic interactions (*i.e.*, hydrogen bonds, charged systems), which are characterized by errors that are twice as large as those found in other regimes.

The origin of this slow convergence arises from two factors. First, only about 20% of the dimers are dominantly bound by electrostatics.[61] The priority of the regression model is thus to minimize the error on the other classes. Second, there is a fundamental dichotomy between the local nature of our symmetry-adapted learning scheme and the long-range nature of the interactions. In fact, the electron density encodes information about the whole chemical system at once, while the machine-learning model represents molecules as a collection of 4 Å wide atom-centered environments. This difference in the spatial reach of the information encoded in the target and in the representation is a limitation. In this respect, a global molecular representation, which includes the whole chemical system, would be more suitable, but this would imply renouncing to the scalability and transferability of the model. Given a large enough training set, however, our SA-GPR model is able to capture the density deformations due to the field generated by the neighboring molecule. The reason is rooted in the intrinsic locality of density deformations and in the concept of “nearsightedness”[253,254] of all local electronic properties, which constitutes a theoretical justification for a local decomposition of such quantities.

The fundamental advantage of setting the electron density as the machine-learning target is the broad spectrum of chemical properties that are directly derivable from $\rho(\boldsymbol{r})$. For instance, the predicted charge densities are the key ingredient in density-dependent scalar fields aimed at visualizing and characterizing interactions between atoms and molecules in real space. Examples of the density overlap region indicator (DORI)[493] are given in Figure 6.4 for representative dimers. Compared to the rather featureless $\rho(\boldsymbol{r})$, DORI reveals fine details of the electronic structure, which constitute a more sensitive probe for the quality of the machine-learning predictions. In particular, it reveals density overlaps (or clashes) associated with bonding and non-covalent regions on equal footing through the behavior of the local wave-vector $(\nabla\rho(\boldsymbol{r})/\rho(\boldsymbol{r}))$.[512–514]

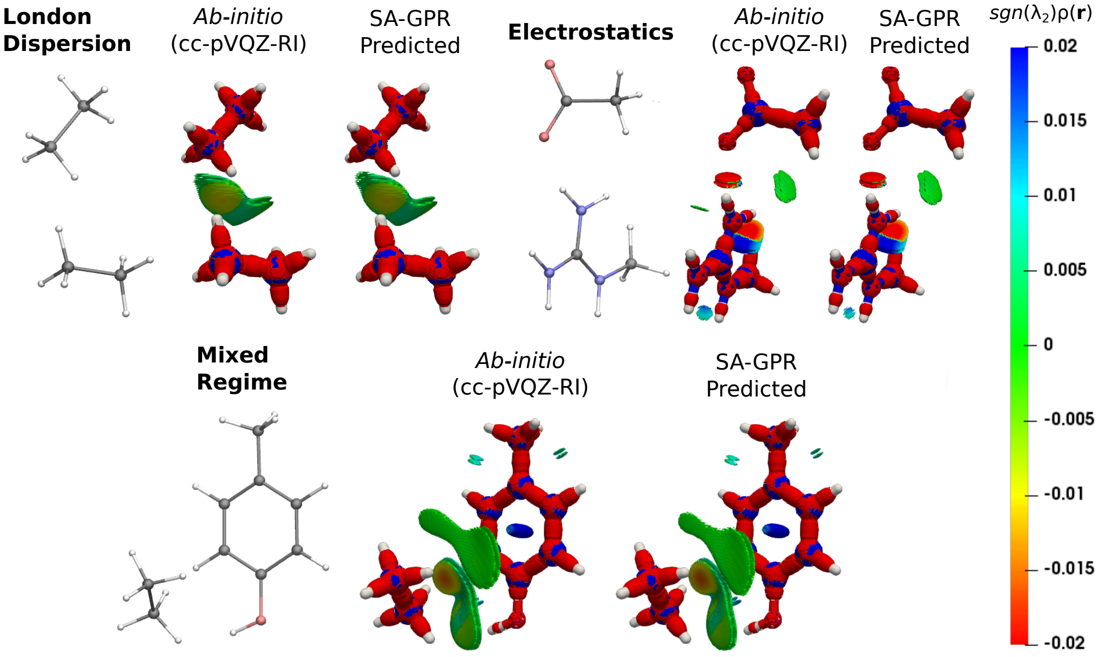


Figure 6.4 – DORI maps of representative dimers for each type of dominant interaction (DORI isovalue: 0.9). Isosurfaces are color-coded[494] with $sgn(\lambda_2)\rho(\boldsymbol{r})$ in the range from attractive -0.02 a.u. (red) to repulsive 0.02 a.u. (blue). In particular, $sgn(\lambda_2)\rho(\boldsymbol{r}) < 0$ characterizes covalent bonds or strongly attractive NCIs (*e.g.* H-bonds); $sgn(\lambda_2)\rho(\boldsymbol{r}) \sim 0$ indicates weak attractive interactions (van der Waals); $sgn(\lambda_2)\rho(\boldsymbol{r}) > 0$ repulsive NCIs (*e.g.* steric clashes).

As shown in Figure 6.4, the intra- and intermolecular DORI domains obtained with the SA-GPR densities are indistinguishable from those in the *ab-initio* maps. This performance is especially impressive for the density clashes associated with low density values, as is typical for the non-covalent domains. All the features are well captured by the predicted densities ranging from large and delocalized basins typical of the van der Waals complexes (in green) to the compact and directional domains typical of electrostatic interactions, to intramolecular steric clashes (*e.g.* phenol, mixed regime). A quantitative measure of the DORI accuracy for the most characteristic basin of each type of interaction is reported in the ESI. Overall, these results illustrate that the residual 0.2% mean absolute percentage error does not significantly affect the density amplitude in the valence and intermolecular regions that are accurately described by the SA-GPR model. The highest amplitude errors are concentrated near the nuclei in the region dominated by the core-density fluctuations.

The versatility of the machine-learning prediction is further illustrated by using the predicted densities to compute the molecular electrostatic potential (ESP) for the same representative dimers (Figure 6.5).

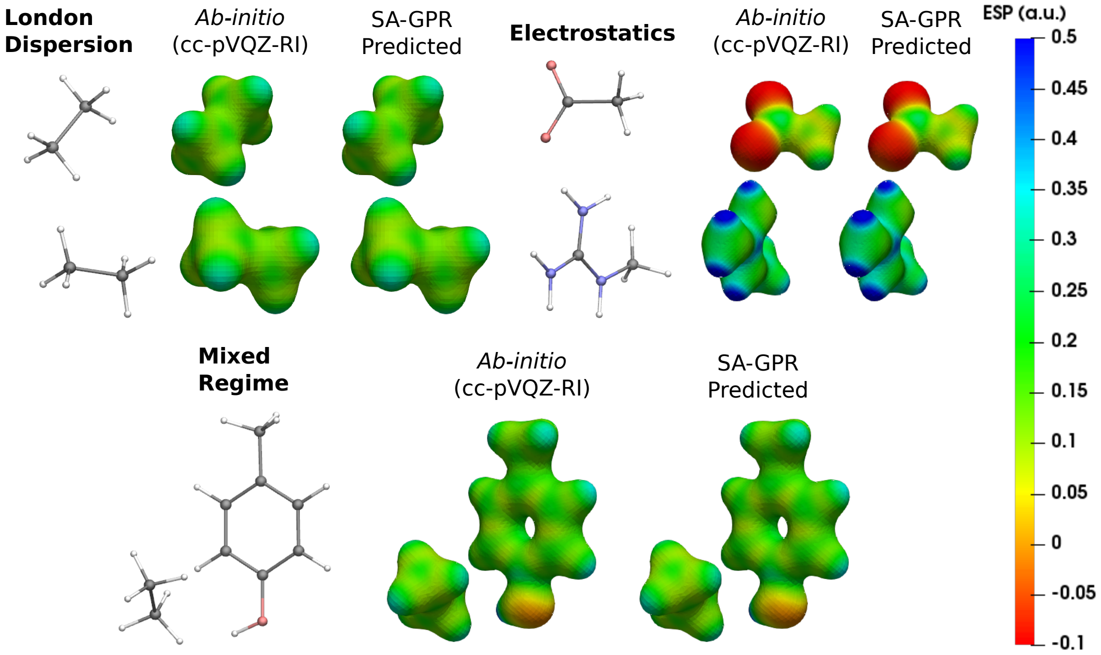


Figure 6.5 – Electrostatic potential (ESP) maps of representative dimers for each type of dominant interaction (density isovalue: 0.05 e^- Bohr^{-3}). ESP potential is given in Hartree atomic units (a.u.).

ESP maps based on predicted densities agree quantitatively with the *ab-initio* reference and correctly attribute the sign and magnitude of the electrostatic potential in all regions of space. Importantly, the accuracy of the ESP magnitude remains largely independent of the dominant interaction type. This is especially relevant for charged dimers (electrostatics) as it demonstrates that despite slower convergence of the learning curve for this category, the achieved accuracy of the model is sufficient to describe the key features of the electrostatic potential.

The most widespread applications of ESP maps exploit qualitative information (*e.g.*, identification of the molecular regions most prone to electrophilic/nucleophilic attack) but the electrostatic potentials can be related to quantitative properties such as the degree of acidity of hydrogen bonds and the magnitude of binding energies.[515–519] As a concrete example related to structure-based drug design, we used a recent model that estimates the strength of the stacking interactions between heterocycles and aromatic amino acid side-chains directly from the ESP maps.[517,518,520] This model derives the stacking energies of drug-like heterocycles from the maximum and mean value of their ESP within a surface delimited by molecular van der Waals volume (at 3.25 Å above the molecular plane).[517] Following this procedure, we used the ESP derived from the ML predicted densities to compute the binding energies between a representative heterocycle included in our dataset, the tryptophan side-chain, and the three aromatic amino acid side-chains (Figure 6.6).

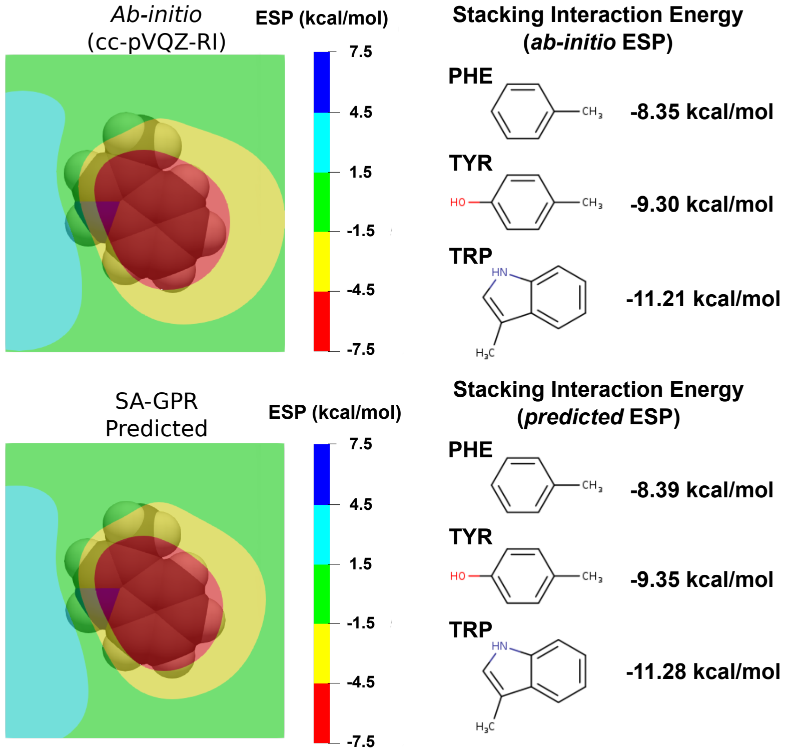


Figure 6.6 – (*left*) Electrostatic potential maps 3.25 Å above the plane of the tryptophan (TRP) side-chain. The van der Waals volume of TRP is represented in transparency. The color code represents the electrostatic potential in kcal/mol according the scale chosen in Ref. 517. (*Right*) Stacking interaction energies of TRP with the phenylalanine (PHE), tyrosin (TYR) and tryptophan (TRP) side-chains computed as detailed in Ref. 517 on the basis of *ab-initio* (*top*) and ML-predicted (*bottom*) ESP.

Comparison between *ab-initio* and ML predicted stacking interaction energies shows that the deviations in the ESP maps lead to minor errors on the order of 0.05 kcal/mol. The largest deviations in the ESP would appear further away from the molecule, beyond the region exploited for the computation of the energy descriptors (*i.e.*, the sum of the atomic van der Waals radii). The predicted ESP shows larger relative deviations far from the nuclei owing to the error propagation of the density predictions $\rho(\boldsymbol{r})$ to the electrostatic potential $\phi(\boldsymbol{r})$. This can be best understood in the reciprocal space, where the deviations of the potential at a given wave-vector **k** are related to the density error by $\delta\hat{\phi}(\boldsymbol{k}) = 4\pi\delta\hat{\rho}(\boldsymbol{k})/k^2$. Because of the k^{-2} scaling, the error on $\phi(\boldsymbol{k})$ increases as $k \to 0$, implying that larger relative errors of the electrostatic potential are expected in regions of space where $\phi(\boldsymbol{r})$ is slowly varying (*i.e.*, thus determined by the long wavelength components).

6.3.1 Prediction on Polypeptides

The tremendous advantage of the atom-centered density decomposition is to deliver a machine-learning model that depends only on the different atomic environments and not on the identity of the molecules included in the training set. Thanks to its transferability, the model provides access to density information of large macromolecules, at the sole price of including sufficient diversity, that can capture the chemical complexity of a larger system. The predictive power of this extrapolation procedure is demonstrated by using the machine-learning model exclusively trained on the 2291 BFDb dimers to predict the electron density of 8 polypeptides taken from the Protein DataBank (PDB).[521] The performance of the ML model for each macromolecules, labelled by their PBD ID, is reported in Figure 6.7.

Overall, the predictions lead to a low average error of only 1.5% for the 8 polypeptides, which is in line with the highest density errors obtained on the BFDb test set. Relevantly, the largest discrepancies are obtained for 3WNE, which is the only cyclopeptide of the set. The origin of these differences can be understood by performing a more detailed analysis on a representative polypeptide, the leu-enkephalin (4OLR). The errors in this percentage range do not affect the density-based properties, such as the spatial analysis of the non-covalent interactions with scalar fields (Figure 6.8 top right panel). Yet, the density differences indicate that the highest absolute errors occur along the amino acid backbone (Figure 6.8 lower panels). In addition, the analysis of the relative error with the Walker-Mezey L(a,a') index[496] shows the highest similarity at the core (99.3%), slowly decreasing while approaching the non-covalent domain (96.3%) (Figure 6.8 top left panel). The L(a,a') index complements the density difference

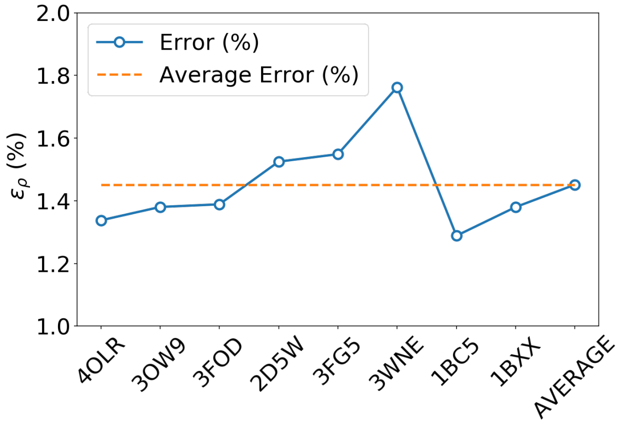


Figure 6.7 – Weighted mean absolute percentage error (ϵ_ρ(%)) with respect to ωB97X-D/cc-pVQZ densities of the predicted densities extrapolated for 8 biologically relevant peptides (protein databank ID).

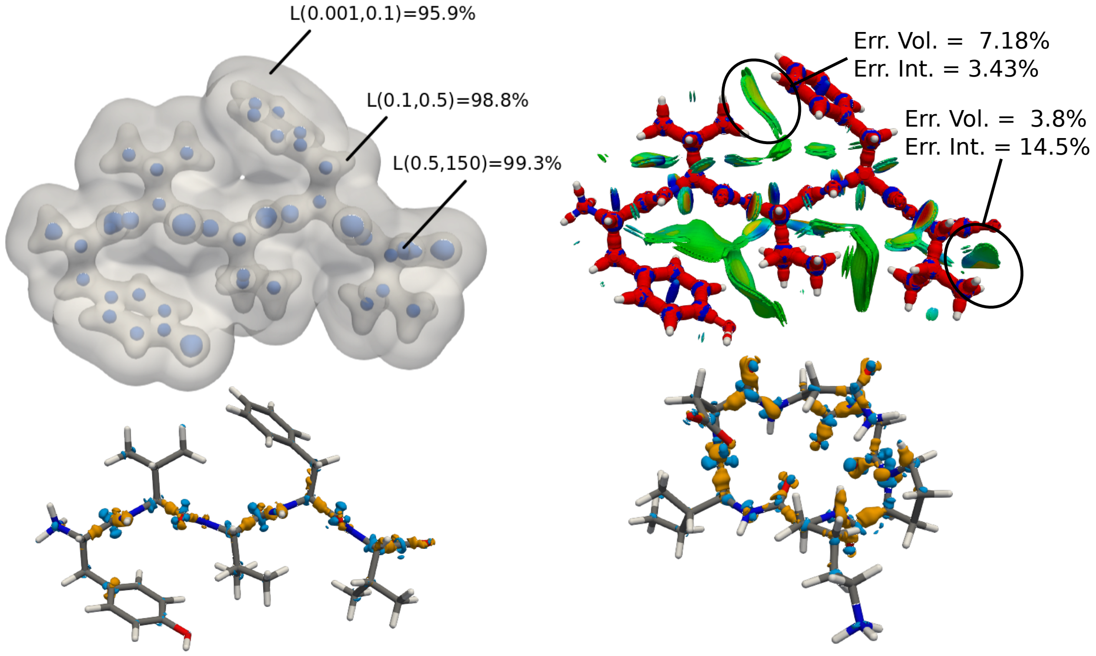


Figure 6.8 – *(top left)* predicted electron density of enkephalin (PBD ID: 4OLR) at three isovalues: 0.5, 0.1, and 0.001 e$^-$ Bohr^{-3}. For each isosurface, the L(a,a') similarity index with respect to *ab-initio* density is reported. *(top right)* DORI map of enkephalin (DORI isovalue: 0.9) colored by $sgn(\lambda_2)\rho(\boldsymbol{r})$ in the range from -0.02 a.u. (red) to 0.02 a.u. (blue) *(lower left)* density difference between predicted and *ab-initio* electron density (isovalues ± 0.01e$^-$ Bohr^{-3}). *(lower right)* density difference between predicted and *ab-initio* electron density of 3WNE (isovalues ± 0.01e$^-$ Bohr^{-3}).

information by showing that the actual density amplitudes and the prediction error do not decrease at the same rate. Nevertheless, the loss of relative accuracy remains modest and the quality of the density is mainly governed by the predictions along the peptide backbone, which are especially sensitive for the more strained 3WNE cyclopeptide. Although similar chemical environments were included in the training set, the error is mainly determined by the lack of an explicit peptide bond motif and cyclopeptides in the training set. While this limitation could be addressed by *ad hoc* modification of the training set, the overall performance of the machine-learning model is rather exceptional as it provides in only a few minutes, instead of almost a day (about 500 times faster for e.g. enkephalin with the same functional and basis set), electron densities of DFT quality for large and complex molecular systems. For comparison, the superposition of atomic densities (*i.e.*, the promolecular approach), which has been used to qualitatively analyze non-covalent interactions in peptides and proteins (*e.g.* Ref. 459) lead to much larger mean absolute percentage errors (17 times higher, see Figure S1 in the ESI).

6.4 Conclusion

Given its central role in electronic structure methods, the total electron density is a very promising target for machine learning, since accurate predictions of $\rho(\mathbf{r})$ give access to all the information needed to characterize a chemical system. Among the many possible properties that can be computed from the electron density, the patterns arising from non-covalent interactions constitute a particular challenge for machine learning models owing to their long-range nature and subtle physical origin. An effective ML model should be transferable across different systems, efficient in learning from relatively small training sets, and accurate in predicting a $\rho(\mathbf{r})$ both in the quickly-varying region around the atomic nuclei, in the tail and – crucially for the study of non-covalent interactions – in those regions that are characterized by low densities and low density-gradients. In this chapter we have presented a model that fulfills all of these requirements, based on an atom-centered decomposition of the density with a quadruple-zeta resolution-of-identity basis set, a symmetry-adapted Gaussian Process regression ML scheme, and training on a diverse database of 2000 sidechain-sidechain dimers extracted from the biofragment database.

The model reaches a 0.3% accuracy on a validation set, that is sufficient to investigate density-based fingerprints of NCIs, and to evaluate the electrostatic potential with sufficient accuracy to quantitatively estimate residue-residue interactions. The transferability of the model is demonstrated by predicting, at a cost that is orders of magnitude smaller than by explicit electronic structure calculations, the electron density for a demonstrative set of oligopeptides, with an accuracy sufficient to reliably visualize bonding patterns and non-covalent domains using the DORI scalar field. Even though the model reaches an impressive accuracy (0.5% mean absolute percentage error) for dimers that are predominantly bound by electrostatic interactions, the comparatively larger error suggests that future work should focus on resolving the dichotomy between the local machine learning framework and the long-range nature of the intermolecular interactions.

7 Learning the energy curvature *versus* particle number

This chapter is based on the following publication:

A. Fabrizio, B. Meyer, C. Corminboeuf, Machine learning models of the energy curvature *versus* particle number for optimal tuning of long-range corrected functionals, *J. Chem. Phys.*, **2020**, *accepted.*

7.1 Introduction

The extension of Hohenberg-Kohn density functional theory (HK-DFT)[56] to non-integer particle numbers led to the determination of two fundamental properties of exact DFT.[13] The first is the piecewise linearity condition, which imposes that the total energy as a function of the (fractional) particle number [*E(N)*] must evolve as a series of straight-line segments.[13,17,62,63] The second is the derivative discontinuity, which establishes that the exact exchange-correlation potential is characterized by sudden jumps while varying across integer particle numbers.[13,522–526]

Approximate density functionals do not fulfill these requirements. Instead, they are generally characterized by a convex *E(N)* curvature and by continuously derivable exchange-correlation potentials.[15,17,65–70] As demonstrated by Kronik, Baer and coworkers,[62] these two quantities are related and therefore the knowledge of the first is sufficient to quantify the extent of the second. Using the same argument, the minimization of the energy curvature has the consequence of correcting the effects of the missing derivative discontinuity, restoring the compliance of approximate functionals to the exact conditions of DFT. Failure to comply with these requirements exacerbates the effects of the delocalization error,[7,17,68,527,528] leads to an incorrect dissociation behavior of heterodimers[13,14,66] and causes the Kohn-Sham frontier orbital eigenvalues to deviate respectively from the ionization potential and the electronic affinity.[77–80]

The existence of a relationship between the curvature and the derivative discontinuity is especially convenient, as the first can be readily evaluated for a given functional and chemical system according to the following expression:[62]

$$C^N_{avg} = \int_{N-1}^{N} C^N(x)\,dx = \epsilon^N_{HOMO} - \epsilon^{N-1}_{LUMO}, \tag{7.1}$$

where C^N_{avg} represent the average curvature between two integer point with N and $N-1$ electrons, while ϵ^N_{HOMO} and ϵ^{N-1}_{LUMO} are the eigenvalues of the frontier orbitals of the N and $N-1$ particle states for a fixed molecular geometry. Equation 7.1 is exact and it is a direct consequence of the Janak's theorem.[77,529]

The straightforward accessibility of the energy curvature information and its relation with fundamental pitfalls of approximate density functionals have been the fertile ground for its use in numerous practical applications. For instance, the minimization of C^N_{avg} serves as a formally motivated criterion for the compound-specific optimal tuning of range-separated hybrid density functionals.[104,105] The accuracy of such functionals has been largely demonstrated in the computations of outer-valence spectra,[530] optical rotations,[531] fundamental and optical gaps.[532–534] The energy curvature has been also applied as a criterion to characterize the severity of the delocalization error in approximate functionals and to rationalize on this basis their relative accuracy.[64,73,86,535–537] In a different context, the curvature information had a central role in the validation of ensemble generalizations of standard density functionals,[63,538–540] carefully designed to retrieve the correct piecewise-linearity behavior of *E(N)* and the derivative discontinuities in the exchange-correlation potential. Finally, the curvature information is applied to develop correction schemes for existing approximate exchange-correlation density functionals.[62,541–544]

The relevance of the information encoded into the energy curvature, corroborated by the extent of its possible applications, contrasts with the modest chemical complexity and relatively low number of molecules for which the curvature has been reported.[62,535]

Recently, machine-learning (ML) techniques have been redefining the scale and the complexity achievable by traditional quantum chemical problems.[545] Supported by the construction of large molecular databases,[24–26,61,245,279,546] the machine-learning approach promotes the large-scale screening of virtually any targeted molecular quantity, with reported examples ranging from simple ground-state properties[41] to complex objects such as electron densities[32–34,55] and the many-body-wavefunction.[35] In addition, machine-learning techniques have been intensively used to promote the access to system-specific quantities such as atomic parameters for semi-empirical computations,[547] atomic and molecular multipoles moments, polarizabilities and overlap integrals in the context of intermolecular potentials.[46,548] Tackling the up-scaling problem with artificial intelligence techniques is especially advantageous, as they reduce the computational cost of accessing molecular properties,[41,549,550] allow extrapolating the acquired information to larger and more complex chemical systems[551,552] and promote the analysis and identification of non-trivial similarity patterns in otherwise unimaginably large amounts of data.[553]

For these reasons, we here report the construction of a machine-learning model of the average energy curvature (C^N_{avg}) of a set of 7165 organic molecules taken from the QM7 database.[24,25] In this chapter, the focus is placed on the curvature between the neutral and the first radical cation state of each molecule, as its minimization leads to compliance with the Koopmans' theorem.[104] The applicability of the regression framework is demonstrated by performing system-specific optimal tuning of the LC-ωPBE functional[67,554,555] based on the predicted curvatures. In addition, the transferability of the model is tested by predicting the optimal range-separation parameters and estimating the ionization potential of two larger molecules of practical use, relevant for the field of hole transport materials. Finally, we address the question of whether specific chemical patterns are more prone to deviation from piecewise linearity using unsupervised dimensionality reduction algorithms to draw statistically robust relationships between the structure/composition of the molecules and their average energy curvature.

7.2 Learning Curves

The training set for the non-linear regression of C^N_{avg} was selected by randomly choosing 6465 molecules out of the QM7 database, leaving the remaining 10% for out-of-sample (oos) predictions. The hyperparameters of the model have been tuned for each functional by 10-fold cross-validation on a randomly selected set containing 10% of the 6465 molecules. The parameters were optimized using a simplex algorithm and an array of 42 different initial conditions sampling a large scale of possible hyperparameter values. At each iteration of the simplex algorithm and for each initial condition, a new set for cross-validation containing 10% of the 6465 molecules was randomly selected. The performance of the model was then evaluated by training on 5 sub-sets of different sizes (100, 500, 1000, 2000 and 5000 molecules) while predicting on a validation set of fixed size (645 molecules). The final learning curve (Figure 7.1) is obtained by randomly sampling the training and the validation set 10 times and averaging the mean absolute errors (10-fold cross-validation). The regression model reported in the Figure uses the spectrum of London and Axilrod-Teller-Muto (SLATM) molecular representation,[250,251] as it was the best performing for the largest training set size (further details in Appendix B).

As shown in the above Figure, the difficulty of the learning exercise largely depends on the level of theory at which the energy curvature is computed. The learning of PBE0[403,404]/def2-SVP and LC-ωPBE[67,554,555]/def2-SVP is the most straightforward, followed by PBE[334,335]/def2-SVP and finally Hartree-Fock (HF/def2-SVP). As already apparent in the upper panels of Figure 7.1, this specific ordering is directly related to the amount of variation of the target quantity (C^N_{avg}) within each functional and method. As shown in Figure 7.2, the mean absolute error of the model trained on 5000 molecules correlates nearly perfectly with the standard deviation of C^N_{avg} for each level of theory.

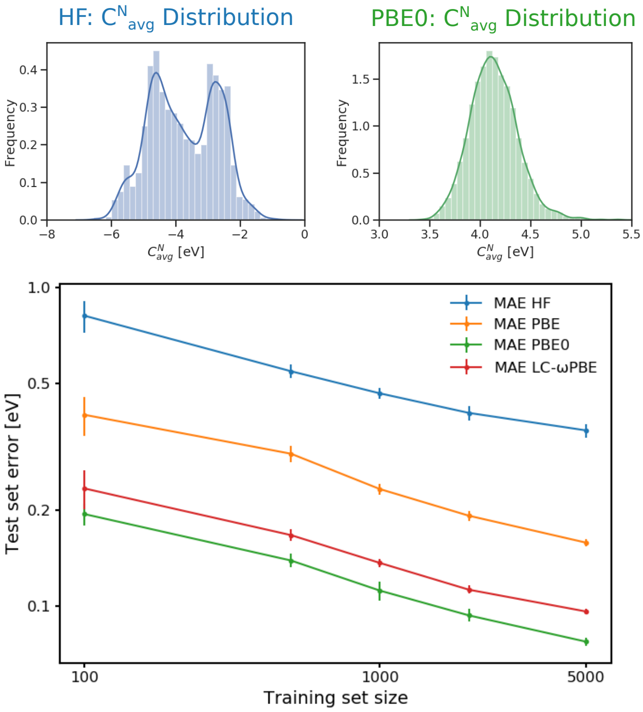


Figure 7.1 – Learning curves of the average energy-curvature (C^N_{avg}) as a function of the training set size. The learning exercise is reported for three functionals and Hartree-Fock (HF) using the def2-SVP basis set. The error bars correspond to the standard deviation of the 10-fold cross-validation. The models were built using the SLATM molecular representation. In the upper panels, we report the distribution of C^N_{avg} across the QM7 database at HF and PBE0.

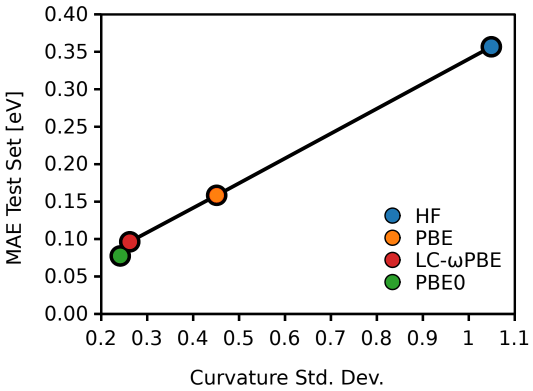


Figure 7.2 – MAE of the model at a training set size of 5000 molecules as a function of the standard deviation of C^N_{avg} using the three functionals and HF.

Following Eq. 7.1, the energy curvature depends on the HOMO eigenvalue of the neutral molecule [N-HOMO] and the LUMO eigenvalue of its radical cation [(N-1)-LUMO]. Therefore, the relative robustness of each functional in describing these two quantities could be invoked to rationalize the overall spread of its C^N_{avg}. However, as shown in the left panel of Figure

7.3, the individual variations of the frontier orbital energies are not sufficient to explain the overall trend found for C^N_{avg}. All the functionals are characterized by similar orbital energies standard deviations, whereas HF shows larger deviations. Importantly, the spread of the individual orbital eigenvalues within each method (Figure 7.3 left panel) is much larger than the standard deviation of their difference (C^N_{avg}, Figure 7.2 x-axis), with the sole exception of HF for which the two are comparable. The narrower distribution of the curvature across the dataset effectively makes this quantity a simpler learning target when compared to previously published efforts to learn individual orbital energies.[28,245,556,557]

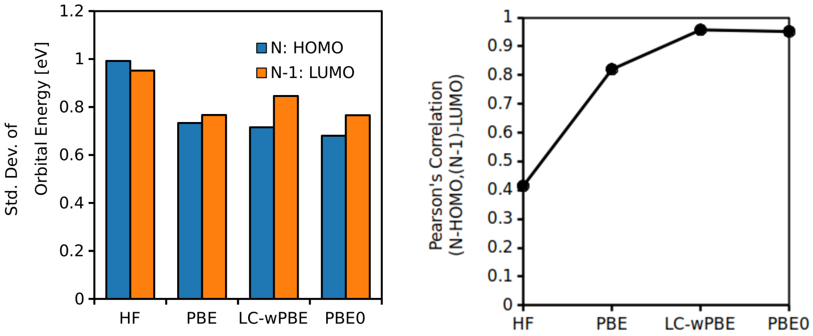


Figure 7.3 – *(left)* Standard deviations of N-HOMO and (N-1)-LUMO through QM7 with three functionals and HF. *(right)* Pearson's correlation coefficient between the N-HOMO and (N-1)-LUMO energies at different levels of theory.

The ordering of Figure 7.2 is retrieved only after combining the information about the variation of the orbital energies with the one about their correlation (Figure 7.3, right panel). In particular, the frontier orbital eigenvalues correlate almost perfectly in LC-ωPBE and PBE0, while their correlation is lower in PBE and very poor within HF. Consequently, the difficulty of the learning exercise ultimately depends on the consistency of a method in describing the orbital energies both of the neutral and the radical cation state of a molecule.

The poor covariance between the frontier orbital eigenvalues in Hartree-Fock is the consequence of the different ways in which the occupied and the unoccupied manifolds are treated within the method. In particular, the orbital energies of the occupied manifold, hence the HOMO eigenvalue of the neutral molecule, is determined in Hartree-Fock by the effective potential of N-1 particles, as the exchange cancels out the self-interaction contribution. This is not the case for the unoccupied manifold, where the effective potential originates from the totality of the particles. In contrast, the energies of both the occupied and the unoccupied orbitals in density functional theory are determined by an N-1 particle effective potential, as the (approximate) exchange-correlation hole excludes a single electron from each and every orbital.

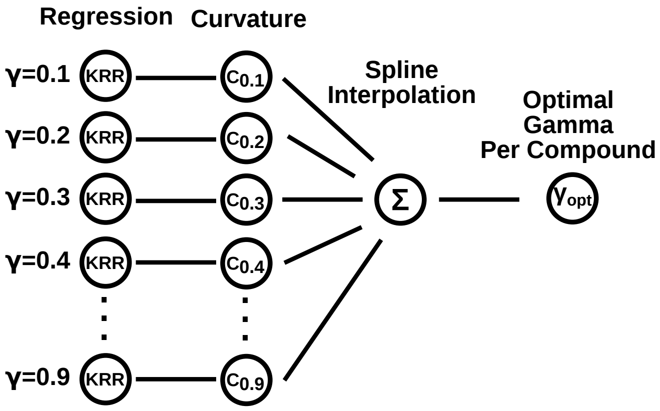


Figure 7.4 – Schematic representation of the regression framework for the prediction of the optimal range separation parameter per compound. For each molecule, nine independent models predict the energy curvature at LC-ωPBE at nine γ values ranging from 0.1 to 0.9 Bohr^{-1}. The system-specific optimal γ parameter, for which $C^N_{avg} = 0$, is then found by a cubic spline interpolation.

7.3 System-specific γ-tuning

The energy curvature predicted for each molecule by the machine-learning model can be readily applied as a criterion for system-specific γ-tuning of range-separated hybrid density functionals. Usually, the tuning procedure consists in adjusting the range-separation parameter to satisfy as closely as possible the Koopmans' theorem for both the neutral and the anionic state of a targeted molecule.[104] This method is by far the most commonly used and relies on the knowledge of the ionization potential either from a computational/experimental reference or from an on-the-fly estimation using ΔSCF procedures.[104,105,558] As already demonstrated by Kronick, Baer, and coworkers[62], the minimization of C^N_{avg} in approximate functionals implies their compliance to the Koopmans' theorem. Therefore, the optimal range-separation parameter for a specific compound can be found by imposing the curvature to be identically zero.

Figure 7.4 schematically illustrates a modification of the regression framework as presented in the previous section, which uses the curvature information to determine the optimal γ parameter for a given chemical system. In particular, the procedure consists of nine independent kernel ridge regression models, each targeting C^N_{avg} at different values of the range-separation parameter. In the last step, a cubic spline interpolation of the predicted curvatures leads to the optimal γ parameter (*i.e.*, the γ value for which $C^N_{avg} = 0$) for a given molecule.

To avoid the introduction of unpredictable noise in the data, we considered here only those compounds, for which all the computations converged. In consequence, the model was trained using the energy curvature of 5754 small organic molecules taken from the QM7 database and used to predict the system-specific optimal LC-ωPBE γ values for a test set of

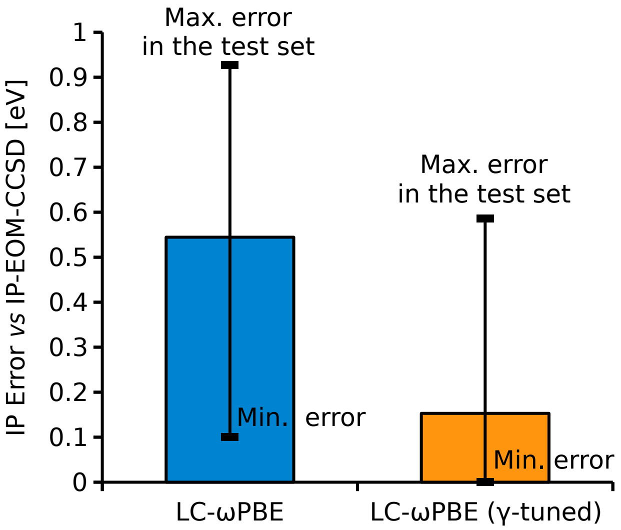


Figure 7.5 – Absolute error between $-\epsilon_{HOMO}$ of LC-ωPBE and its γ-tuned variant and the ionization potential at IP-EOM-CCSD across the 640 molecules of the test set. Optimal γ values derive from the model described in Figure 7.4. The height of the histogram represents the mean absolute error, while the bars show the maximum and the minimum deviations.

640 molecules. Upon a single point computation using the tuned functional, the ionization potential of each molecule is evaluated as $-\epsilon_{HOMO}$ and compared to the corresponding value at IP-EOM-CCSD. For consistency, all computations are performed with the def2-SVP basis set. Figure 7.5 shows the accuracy of estimated IPs averaged over the test set for the standard LC-ωPBE and its γ-tuned variant. The error bars show the maximum and the minimum deviation from the IP-EOM-CCSD reference registered among the 640 molecules.

The tuning procedure based on the predicted curvatures results in a five-fold decrease of the average ionization potential error compared to the standard functional. The robustness of the predictions is further demonstrated by the fact that the worst error made with the γ-tuned variant is only as high as the average error made with the standard LC-ωPBE.

Including several hundreds of different molecules, the test set represents a sufficiently large ensemble for a statistically relevant analysis of the optimal range-separation parameter in LC-ωPBE. By registering the frequency of appearance of the predicted γ values, it is shown that their distribution tends to a Gaussian function centered around 0.32 Bohr^{-1} (Figure 7.6). Out of the 640 molecules, only 12 are characterized by an optimal γ parameter close to the 0.4 Bohr^{-1} of the standard functional. In all those cases where system-specific γ-tuning is not possible, for instance in the computation of dimer binding energies,[106] the distribution in Figure 7.6 demonstrates that fixing the range-separation parameter of LC-ωPBE to 0.32 Bohr^{-1} would reduce the curvature for the majority of molecules.

The discrepancy with the original parametrization of LC-ωPBE has to be interpreted as the results of a different optimization strategy. Here, the suggested 0.32 Bohr^{-1} minimizes the

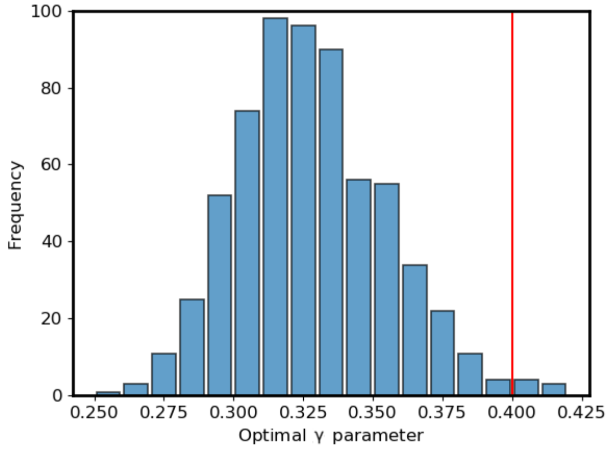


Figure 7.6 – Distribution of the optimal γ parameters [Bohr^{-1}] across the 640 molecules of the test set as predicted by the model described in Figure 7.4. The red line show the value of the range-separation parameter in the standard LC-ωPBE.

energy curvature for the highest number of compounds in a comprehensive dataset of organic molecules. Following the works of Baer,[62,104] fixing the γ parameter by minimization of the energy curvature is a formally motivated procedure, as it leads to compliance with the Koopmans' theorem and exact conditions of DFT. The original approach used for the parametrization of LC-ωPBE is more pragmatic and seeks to minimize the error of the functional against different energy-based benchmark databases.[555] The formal issue associated with this second strategy is that the range-separation parameter inevitably compensates for unrelated deficiencies in the rest of the approximated exchange-correlation functional.

7.4 Extrapolation

The machine-learning models presented in the previous paragraphs rely on a global molecular representation, *i.e.* each vector in the feature space characterizes one specific compound. As the energy curvature is a molecular property, this kind of representation is highly suitable and easily applicable to the regression problem. On the other hand, a model based on a global molecular representation is not transferable: it cannot be trained on simple compounds and used to predict larger molecules.[559] This issue can be tackled using local representations, which encode the molecular information as a collection of atoms in their environments. By establishing similarity measures between local atomic environments, rather than between whole molecules, local representations lead to transferable models, applicable to larger and more diverse molecules than those included in the training set (see, for instance, Refs. [34,560, 561]). The regression framework shown in Figure 7.4 is general and can be readily extended to local, atom-centered molecular representations. More details about the modification of the learning framework to accommodate locality and transferability are given in Appendix B.

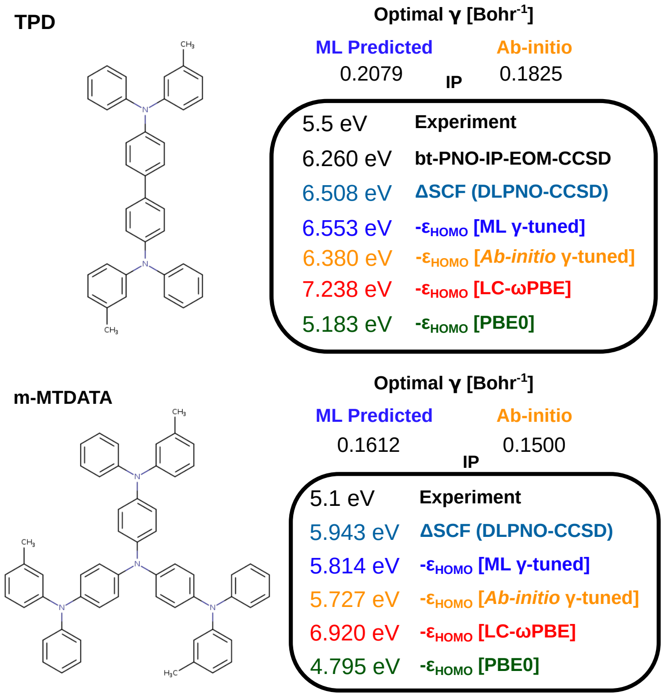

TPD
Optimal γ [Bohr^{-1}]
ML Predicted
0.2079
Ab-initio
0.1825
IP
5.5 eV Experiment
6.260 eV bt-PNO-IP-EOM-CCSD
6.508 eV ΔSCF (DLPNO-CCSD)
6.553 eV -ε$_{HOMO}$ [ML γ-tuned]
6.380 eV -ε$_{HOMO}$ [*Ab-initio* γ-tuned]
7.238 eV -ε$_{HOMO}$ [LC-ωPBE]
5.183 eV -ε$_{HOMO}$ [PBE0]
m-MTDATA
Optimal γ [Bohr^{-1}]
ML Predicted
0.1612
Ab-initio
0.1500
IP
5.1 eV Experiment
5.943 eV ΔSCF (DLPNO-CCSD)
5.814 eV -ε$_{HOMO}$ [ML γ-tuned]
5.727 eV -ε$_{HOMO}$ [*Ab-initio* γ-tuned]
6.920 eV -ε$_{HOMO}$ [LC-ωPBE]
4.795 eV -ε$_{HOMO}$ [PBE0]

Figure 7.7 – Extrapolation: optimal γ parameter derived from the model described in Figure 7.4 for two large molecules relevant for the field of hole-transporting materials compared to the value obtained by *ab-initio* optimal tuning. The value of $-\epsilon_{HOMO}$ for both the standard LC-ωPBE and its γ-tuned variant (ML and *ab-initio*) are reported along with reference ionization potentials. Experimental IPs are taken from Refs. [562,563]

Figure 7.7 shows the application of the local regression framework to predict the optimal γ values of two large molecules commonly used in hole-transporting materials:[564,565] N,N'-Bis(3-methylphenyl)-N,N'-diphenylbenzidine (**TPD**) and 4,4',4"-Tris[(3-methylphenyl)phenylamino] triphenylamine (**m-MTDATA**). The model was exclusively trained on the local environments of the small organic molecules of the QM7 database using the atomic spectrum of London and Axilrod-Teller-Muto (aSLATM)[250,251] representation.

The $-\epsilon_{HOMO}$ computed with standard LC-ωPBE is a rather poor approximation of the ionization potential of **TPD** and **m-MTDATA** with errors around 1 eV compared to the *ab-initio* references (bt-PNO-IP-EOM-CCSD and ΔSCF at DLPNO-CCSD). Upon ML-based γ-tuning the error with respect the wavefunction based methods is reduced to 0.1-0.2 eV for both molecules. In addition, the optimal γs from machine learning and the corresponding IPs are in very close agreement with the optimal parameters obtained *ab-initio* by minimizing the difference between $-\epsilon_{HOMO}$ and the neutral-cation ΔSCF energy ($\Delta\gamma = 0.025$ Bohr^{-1} and 0.01 Bohr^{-1}). These results, obtained on compounds four times larger than the largest molecule in the training set, demonstrate the transferability of the local model and its applicability to targeted complex molecules. Interestingly, the optimal γ parameters for both **TPD** and **m-MTDATA** are much lower than any value obtained on the smaller molecules of the QM7 test set (Figure 7.6). This behavior is consistent with the results of the existing literature[85,566–568] and further supports the conclusion that γ can be interpreted as the inverse of an effective conjugation length dependent on the system size. Finally, the HOMO eigenvalue of PBE0 is the farthest from the *ab-initio* reference, but the closest to the experimental values obtained by cyclic voltammetry in organic solution (TPD)[562] or by ultraviolet photoemission spectroscopy in amorphous solid-state (m-MTDATA).[563] This result is not unexpected (see, for instance, Refs. 569,570) and shows that the error made by the global hybrid mimics the effects of the condensed phase environment (*e.g.* solvent, crystal field).[571,572]

7.5 Unsupervised learning and analysis of the QM7 dataset

The large chemical diversity contained in the QM7 database promotes a thorough assessment of the relation between the energy-curvature computed with a given functional and the system-specific structural and compositional patterns. However, drawing such a relationship for thousands of molecules inevitably leads to a high-dimensional problem, which is unsuitable for analysis and visualization. In this context, non-linear dimensionality reduction algorithms reveal the underlying structure of high-dimensional data by projecting complex vectors into lower dimensions. Figure 7.8 shows a two-dimensional representation of the chemical diversity of the database using t-distributed Stochastic Neighbor Embedding (t-SNE).[573] This algorithm converts the similarity between molecules, which is defined herein as the euclidean distance between their SLATM representation, to the probability of being each other's neighbors. The embedding of high-dimensional data into lower dimensions is then performed by ensuring that the joint probability between molecules should not change upon projection. While the two axes (dimensions) obtained after a t-SNE transformation have no formal physical or chemical meaning, it is still possible to identify at least a qualitative correlation between chemical properties and the dimensions in Figure 7.8 *vide infra*.

The application of t-SNE to QM7 on the basis of the SLATM representation for each molecule SLATM reveals clusters of compounds with similar chemical patterns, mainly defined by the presence or the absence of heteroatoms and their connectivity. In particular, the vertical axis (Dim. 2) somehow correlates with the number of heteroatoms, from zero (alkanes, bottom) to two or more non-carbon atoms (hydroxyamines and oxyamines, top). The horizontal axis follows instead a gradient of chemical composition going from the oxygen-based compounds (left) to nitrogen-containing molecules (right), passing from mixed species. Each point is color-coded by its average energy curvature computed at PBE/def2-SVP to establish a global, qualitative connection between these macro-families of compounds and the degree of their deviation from piecewise linearity. The choice of PBE is motivated by the fact that the absence of Hartree-Fock exchange leads to a curvature that represents an upper limit for the other functionals.

Figure 7.8 highlights seven key families characterized by at least one region of high average energy curvature (in red). Out of those clusters, three contains only oxygen as heteroatom

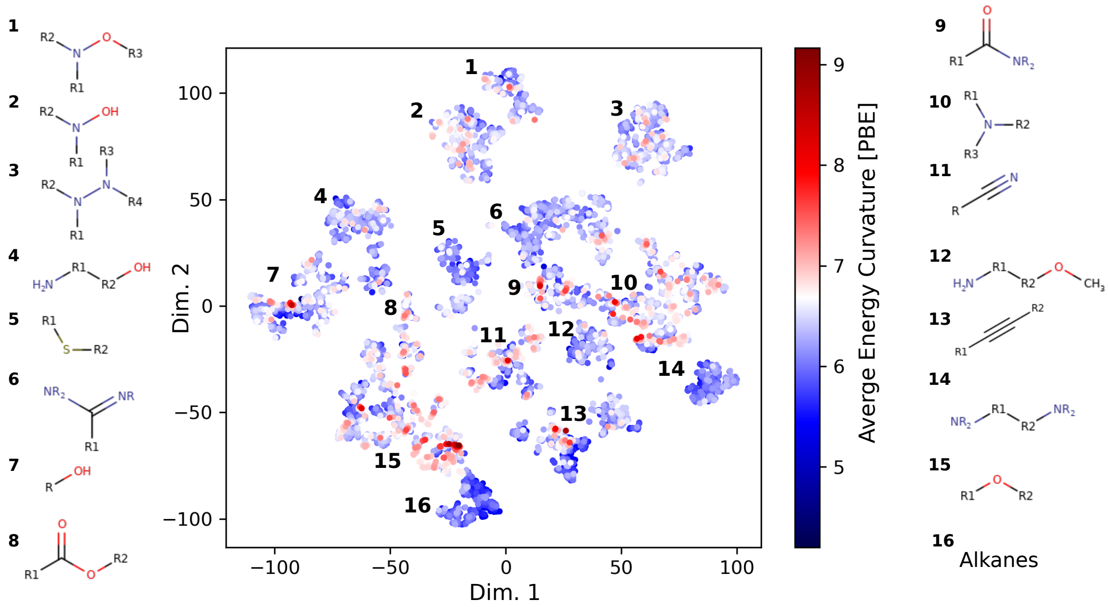


Figure 7.8 – Two-dimensional t-SNE map of the QM7 database on the basis of the SLATM representation. Each point represents a compound colored by its average energy curvature computed with PBE/def2-SVP. The diverging color map highlights the data with the highest and the lowest average energy curvature. Each of the clusters contains molecules with similar patterns, which are defined by the corresponding numbering.

(alcohols[7], ethers[15] and acids/esters [8]), two includes *sp*-hybridized carbons (cyano groups [11] and alkynes [13]), one contains only nitrogen (amines [10]) and the last group includes the amides [9]. In contrast, alkanes (with the exception of the smallest methane and ethane, see discussion below)[16], diamines separated by long carbon chains [14], all sulfur-containing compounds [5] and amidines [6] are all characterized by lower curvatures. These trends suggest that the presence of increasingly electron-rich heteroatoms tends to increase the average energy curvature. In particular, the presence of oxygen atoms is especially sensitive as shown by the qualitative difference between amides and amidines. The low average energy curvature that characterizes all sulfur-containing compounds suggests that the presence of heteroatoms beyond the second row of the periodic table does not have a critical impact on the deviation from piecewise linearity. In addition to heteroatoms, the hybridization of the carbon centers is also a relevant factor as illustrated by the contrast between alkanes and alkynes groups. These conclusions are consistent with previous work on charge transfer complexes[574,575] and delocalization error.[537] In particular, the results presented here are comparable with the work of Kronik and Baer,[62] who report the average energy curvature for a set of nine small molecules, whose order can be rationalized in terms of the presence of electron-rich heteroatoms, their hybridization, and the molecular size.

Although not explicitly evident in the mapping of Figure 7.8, the molecular size is, in fact, crucial to determine the extent of the average energy curvature. To emphasize this point,

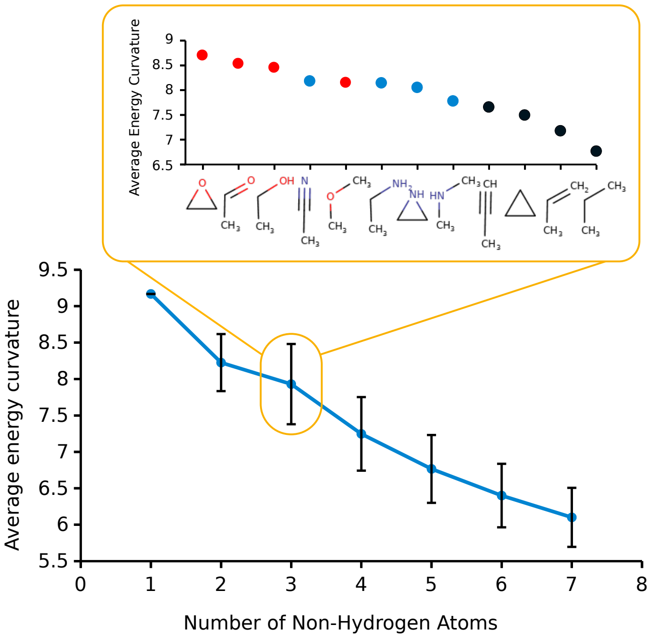


Figure 7.9 – Correlation between the average energy curvature at PBE/def2-SVP and the size of the molecules. The mean values (dots) are computed averaging C^N_{avg} over all the compounds with the same number of non-hydrogen atoms. The error bar represents one standard deviation from the mean. The inset shows the average energy curvature of all the compounds in QM7 with 3 non-hydrogen atoms. The color code in the inset highlights the presence of oxygen (red), nitrogen (blue) or carbon only compounds (black).

Figure 7.9 correlates the curvature at PBE/def2-SVP and the size of the molecules, upon averaging C^N_{avg} over all the molecules with the same number of non-hydrogen atoms (N_{heavy}). Although the mean values for $N_{heavy} = 1$ and $N_{heavy} = 2$ are not statistically significant (*i.e.*, these categories include only 1 and 3 molecules respectively), the robust inverse size/curvature relationship justifies the high curvature of the smallest alkanes (Cluster 16 Figure 7.8).

The error bars in Figure 7.9 shows that within every N_{heavy} there is a distribution of curvatures that reflect the chemical composition. The analysis of the subset with 3 non-hydrogen atoms is especially suitable as it contains sufficient compounds to reflect general trends but is simultaneously small enough to list all its molecules. The inset of Figure 7.9 shows the energy curvature of all the compounds with $N_{heavy} = 3$ ordered from the highest to the lowest. This plot validates the conclusions drawn from the t-SNE map, as the curvature decreases with the electron-richness of the heteroatom (O > N > C). One exception due to the effects of hybridization is acetonitrile, which has a slightly higher, but comparable curvature to methyl ether. Complementing the information of the t-SNE map, the inset Figure reveals the high-energy curvature of 3-membered rings (oxirane, aziridine, and cyclopropane) that are generally considered to act as unsaturated systems.[576]

7.6 Conclusion

The average energy curvature with respect to the particle number is a crucial system-dependent property of density functionals, which quantifies their deviation from the exact conditions of DFT and therefore affects their accuracy. Related to the lack of derivative discontinuity in the exchange-correlation potential and thus to the degree of severity of the delocalization error, the information about this quantity has been successfully used for optimal tuning of long-range corrected functionals and to correct Kohn-Sham orbital eigenvalues to match ionization potentials and electron affinities. In this chapter, we have proposed the construction of a machine-learning model of the average energy curvature and shown its applications for the system-specific tuning of the LC-ωPBE functional. In parallel, unsupervised learning techniques have been applied to obtain qualitative information about particular chemical patterns and molecular properties which results in highly convex curvatures.

As the curvature is both a system-specific and a functional dependent quantity, we have first shown that the learning exercise is not equally difficult for any given functional, but it depends on its ability to describe on equal footing the neutral and the radical cation state of a molecule. This result implies that the possible spread of value for the average energy curvature is not equal for all methods. In particular, the largest standard deviation for the curvature is registered for Hartree-Fock, due to the poor correlation between the neutral molecule HOMO eigenvalue and the LUMO of the radical cation.

Training several independent models to target the curvature at LC-ωPBE for different values of its range-separation parameter led to the construction of a second framework dedicated to the system-dependent optimal tuning of the functional. The use of the predicted γ parameters resulted in a five-fold increase of the accuracy when estimating the first ionization potential (IP) with $-\epsilon_{HOMO}$ with respect to the standard functional. The distribution of the predicted range-separation parameters on the QM7 database shows that the original 0.4 value of LC-ωPBE is far from optimal to minimize energy curvature. As a generalization of the framework, we use a local molecular representation for the training and demonstrate the transferability of the modified model by estimating the optimal γ-values and computing the ionization potentials of two larger molecules, relevant for the field of hole-transporting materials.

Finally, projecting the high dimensional SLATM representation of QM7 in two dimensions with a t-SNE algorithm revealed the underlying structure of the database. In particular, the mapping showed several distinct clusters enclosing molecules similar to each other in terms of their scaffold and presence of heteroatoms. The curvature values across these clusters were found to assume the highest values for compounds with second row heteroatoms, most frequently oxygen, or for compounds with sp-hybridization. Additional analysis of the data supports the existence of an inverse correlation between molecular size and the average energy curvature.

7.7 Computational Details

The molecular geometries for all species were taken as published in the QM7 database.[24,25] The curvatures were computed according to Equation 7.1 using the orbital eigenvalues of the neutral and the first radical cation state of each molecule. All the computations using PBE, PBE0, LC-ωPBE and Hartree-Fock were performed in Gaussian16,[305] in combination with the def2-SVP[325] basis set. The first ionization potential energies at IP-EOM-CCSD[577,578] and bt-PNO-IP-EOM-CCSD[579,580] were obtained with Orca 4.0[318] using the def2-SVP basis set for consistency with the DFT values. The density fitting approximation was applied in the bt-PNO-IP-EOM-CCSD computation. The machine-learning representations and similarity kernels were obtained using the Quantum Machine Learning toolkit QMLcode[581] with the exception of SOAP[59] (see Appendix B), which was computed using DScribe 0.3.2.[582] The mathematical form of the similarity kernels was chosen as standard procedure according to the specific representation. The two-dimensional map of the QM7 database was generated using the t-SNE[573] algorithm as implemented in the scikit-learn package.[583]

8 General Conclusions and Outlook

Modern quantum chemistry stands at a thrilling stage of its development. Along with the challenges of the traditional approaches, recent technological innovations such as GPU accelerated software and machine learning are redefining the paradigms of the field. The work presented in this thesis reflects this evolution and simultaneously demonstrates that the deterministic and the statistical perspectives to quantum chemistry are complementary. In this context, the role of density functional theory is central as it provides a common underlying framework for all the material covered. The chemical situations, the methods and the applications discussed here belong to a larger effort that aims at broadening the domain of applicability of density functional theory and are part of a more comprehensive collection of novel methodologies to access fundamental chemical properties of complex molecular systems.

The first part of this thesis focused on identifying concrete chemical situations, which still represent a challenge for the application of standard computational procedures, approximate density functionals, and correction schemes. Built in 2012, the Orel26rad dataset has already demonstrated the failure of common density functionals in computing the interaction energies of simple radical cation dimers, which are the fundamental functional units in organic electronic materials. With the construction of CryOrel9, the conclusions drawn from Orel26rad are extended to a more realistic set of radical cation dimers, taken from the crystal structure of actual organic semiconductors and representative of their different crystal arrangements. Using both datasets, we analyzed the performance of recent functionals of the ωB97X series and proposed a variant of ωB97X jointly-fitted with the dDsC density-dependent dispersion correction. The ability to balance delocalization error and London dispersion interactions, capital in Orel26rad, is even more crucial in CryOrel9 as both these effects grow with the system size, but decay at a different pace. The use of density overlaps in the ωB97X-dDsC damping function improves the description of medium-range interactions, making the functional particularly suitable to compute the properties of radical cation dimers. The combinatorially-optimized ωB97M-V yields highly accurate results for some radical cationic dimers, but its robustness across chemical diversity and different spatial arrangements is not guaranteed. Results from U-SAPT0 analysis on CryOrel9 revealed that not all the supramolecular arrangements are

equally challenging. Planar, π-stacked dimers are the most difficult systems to describe, limiting the overall accuracy of most of the density functionals, including ωB97X-D, ωB97X-D3 and ωB97X-V.
The second example reported focuses on the modulation of excited state processes by London dispersion interactions in molecular switches bearing bulky lateral substituents. In particular, we have compared the static excited state energy profiles and the outcome of excited stated molecular dynamics simulations of a prototypical photoswitch (stilbene) and a substituted variant (3,3',5,5'-tetra-*tert*-butyl-stilbene) using dispersion corrected functionals, their uncorrected variants, and CC2 as a reference method. These computations revealed that while the rearrangement of the electronic density upon excitation remains the principal driving force of the excited state processes, the inclusion of London dispersion interactions is crucial to correctly describe the potential energy surfaces and thus the structural evolution and photo-deactivation pathways of the excited molecules. Failure to account for van der Waals interactions leads to qualitatively incorrect results for the substituted stilbene, and in particular to a spurious hindrance of its photocyclization pathway. Overall, London dispersion interactions beyond the common ground-state chemical situation cannot be neglected *a priori*. In addition, we have demonstrated that standard dispersion corrections, parametrized in principles only for the ground-state, are nevertheless largely beneficial to describe the excited state processes of molecules with sizable, but otherwise not photoexcited substituents.

Alongside the challenge of improving existing deterministic approaches as discussed in the previous paragraph, the recent advances in the machine learning technology defy quantum chemistry to provide learnable representations of complex, yet fundamental, molecular properties. In this context, the second part of the thesis reports the construction, the refinement and the applications of a local and transferable machine learning model of the electron density. Inspired by well-established linear-scaling, embedding and fragmentation methods, we proposed a decomposition Ansatz for the electron density on atom-centered, non-orthogonal basis functions. This local representation of $\rho(\boldsymbol{r})$ provides a suitable target for symmetry-adapted Gaussian process regression, as demonstrated with the accurate prediction of the valence electron density for a set of hydrocarbons of increasing complexity. In addition, the transferability of the model is shown by predicting the electron density of octane and octatetraene, while only training on butane and butadiene.
This first, proof-of-principles, model is then refined using a specialized density-fitting basis set for the decomposition. This choice is shown to be two-fold beneficial. First, the residual basis set error is halved with respect to standard basis sets, which are optimized to model the wavefunction. Second, within the density-fitting framework, the basis set expansion coefficients can be computed analytically, simplifying the regression of the core electron density. The improved model is tested on a challenging set of amino acid side-chains dimers, taken from the Biofragment database (BFDb). The accuracy of the predicted electron densities has been shown for a variety of applications, including the identification of covalent and non-covalent interaction fingerprints with the DORI scalar field, the computation of electrostatic potentials and the estimation of stacking interaction energies of planar heterocycles. The transferability

is further demonstrated with the prediction of the electron density of a series of complex polypeptides.

Finally, Kohn-Sham density functional theory and machine learning are combined in the last work. Here kernel-based non-linear regression and dimensionality reduction algorithms are applied on a comprehensive database of small organic molecules (QM7) to analyze and correct one of the well-known limitations of approximate density functionals: the spurious energy *versus* particle number curvature. In particular, the focus has been placed on the average energy-curvature between the neutral and the first radical cation state, which is responsible for the deviation of the Kohn-Sham HOMO eigenvalue from the first ionization potential. As a first result, we show that learning the average energy-curvature is not equally difficult for any given functional, but it depends on its ability to describe on equal footing the neutral and the radical cation state of a molecule. The information obtained from the trained models is then used to develop a second framework for the γ-tuning of LC-ωPBE. Using the predicted range-separation parameters allows a five-fold increase in the accuracy of the first ionization potential computed as the negative of the Kohn-Sham HOMO eigenvalue. Reference IPs have been evaluated at IP-EOM-CCSD level. The distribution of the predicted range-separation parameters on the QM7 database shows that the original 0.4 value of LC-ωPBE is far from optimal. Instead, the distribution of the predicted parameters is characterized by an expectation value of 0.32. As a final result, we apply t-distributed Stochastic Neighbor Embedding (t-SNE) as a dimensionality reduction algorithm to highlight specific chemical patterns that are prone to large energy-curvature.

The results presented in the previous paragraphs summarize the main conclusions of this work and set the stage for new compelling perspectives, as outlined in the following sections:

Extension of the machine learning model of the electron density to condensed phase and excited states.

The machine learning model of the electron density reported in Chapter 5 and 6 represent a particular case of a much more general framework, applicable to many different chemical situations. Its most simple and readily available generalization consists in the regression of a state-specific density, *e.g.* the response density from TDDFT computations, instead of the ground-state density only. Slightly more involved from the quantum chemical perspective is the regression of the transition density, as it would require the reformulation of the local decomposition procedure. On the other hand, the ability to compute transition densities at a fraction of the actual *ab-initio* cost would be invaluable to address large scale excited state problems, such as the combinatorial screening of thousands of photochemically active molecules and their properties.

Besides targeting systems beyond their electronic ground-state, a different extension of the model could focus on the condensed phase and in particular, on molecular crystals. This goal could be achieved by combining the existing machine learning architecture with the Gaussian and augmented-plane-wave method, which allows expressing the density matrix of periodic systems in terms of an atom-centered, local basis. This framework would allow translating to the condensed phase all the analysis tools presented in the previous Chapters, with broad applications ranging from screening to the identification of polymorph fingerprints in molecular dynamics simulations.

Drawing a robust statistical relationship between $\rho(\boldsymbol{r})$ and the exchange-correlation energy with machine learning.
As first highlighted by Becke,[342] the guaranteed existence and uniqueness of a functional relationship between the energy and the electron density fundamentally justify the introduction of a certain degree of empiricism in the development of approximate density functionals. Taking this premise to the extreme, most of the physical arguments used in the construction of exchange-correlation density functionals could be discarded, in favor of a purely statically drawn relationship between $\mathrm{E_{XC}}$ and $\rho(\boldsymbol{r})$. In perspective, this could be achieved by constructing a non-linear regression model of the exchange-correlation hole ($h_{\mathrm{XC}}(\mathbf{r},\mathbf{r}')$), which defines the $\mathrm{E_{XC}}$ as follows,

$$E_{\mathrm{XC}}[\rho] = \frac{1}{2}\int d\mathbf{r}\int d\mathbf{r}'\rho(\mathbf{r})\frac{h_{\mathrm{XC}}(\mathbf{r},\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\rho(\mathbf{r}'). \tag{8.1}$$

Targeting $h_{\mathrm{XC}}(\mathbf{r},\mathbf{r}')$ would result in a truly universal model, as the system-dependence of $\mathrm{E_{XC}}$ would be exclusively encoded in the density. The construction of such a model presents, nevertheless, difficulties of a different nature. From the quantum chemical perspective, the challenge consists in finding a form of Equation 8.1 compatible with the learning framework. On the other hand, the machine learning model has to be carefully constructed to include all the symmetries of the exchange-correlation hole and to avoid numerical instabilities related to the divergence of the Coulomb potential for $\boldsymbol{r} = \boldsymbol{r}'$.

Tackling the static correlation problem with symmetry-adapted machine learning: the on-top pair density.
The on-top pair density $[\Pi(\boldsymbol{r})]$ is a local electronic property defined as the probability of finding two electrons at the same position in space:[22]

$$\Pi(\boldsymbol{r}) = \binom{N}{2}\sum_{\sigma,\sigma'}\int|\Psi(\boldsymbol{r},\sigma,\boldsymbol{r},\sigma',\boldsymbol{x}_3,...,\boldsymbol{x}_N)|^2 d\boldsymbol{x}_3...d\boldsymbol{x}_N \tag{8.2}$$

As originally proposed by Becke, Savin and Stoll,[584] standard Kohn-Sham density functional theory can be generalized to describe multideterminental states by reformulating the exchange-correlation functionals in terms of the total electron density and the on-top pair density. This early work has paved the way for the more recent development of multi-configuration pair-density functional theory[585–593] and the use of $\Pi(\boldsymbol{r})$ as an effective metric to quantify static correlation in simple molecular systems.[594] However, obtaining accurate on-top-pair densities can be a computationally demanding task especially for large chemical systems. Therefore, the construction of a transferable machine learning model of $\Pi(\boldsymbol{r})$ is an appealing perspective, which would set the stage to address the static correlation problem in Kohn-Sham DFT bypassing *ab-initio* computations. The existing symmetry-adapted GPR architecture could be used to attain this objective, on the condition of proposing an efficient local decomposition scheme able to capture the radial and the angular features of the on-top pair density.

A CryOrel9 and ωB97X-dDsC

A.1 Functional performance on CryOrel.

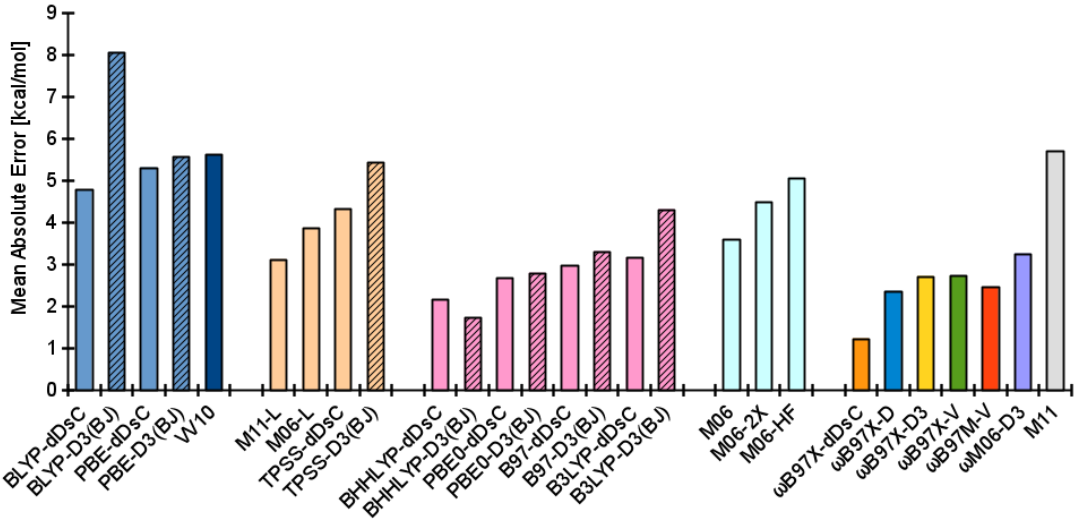


Figure A.1 – MAE of tested functionals on the CryOrel dataset. Blue: GGAs; pale orange: meta-GGAs; pink: global hybrids; cyan: meta-hybrids. Range-separated hybrids follow the same color code as in previous Figures. Oblique bars are drawn to mark the same functional used with a different dispersion correction.

As visible in Figure A.1, the robustness of the theoretical method employed for describing the radical cationic dimers is dictated by its ability to treat delocalization error and London dispersion interactions. These two aspects have been tested using an illustrative set of GGA (B97,[342] PBE[334,335] and BLYP[303,595]), meta-GGA (TPSS,[326] M06L[366] and M11L[596]), global hybrid (B3LYP,[82,303,304] PBE0[403,404] and BHHLYP[81,303]), meta-hybrid (M06,[597] M06-2X[597] and M06-HF[132]) and range-separated hybrid (ωB97X-D,[155] ωB97X-D3,[299] ωB97X-V,[300] ωB97M-V,[280] ωM06-D3,[299] M11[598], ωB97X-dDsC) functionals. Long-range dispersion interactions were accounted for using *ad hoc* post-SCF dispersion corrections (-D3(BJ)[160] and

-dDsC[169–171]), as well as nonlocal correlation functionals (*i.e.*, VV10,[137] for ωB97X-V, ωB97M-V) or effective one-electron potentials (M06 family).

A.2 ωB97X-dDsC

Equations A.1, A.2 and A.3 illustrate the general form of the GGA exchange-correlation core, which is characteristic of the B97 family.[342]

$$E_X^{GGA} = \sum_\sigma \int d\boldsymbol{r} \cdot e_{X\sigma}^{LSDA}(\rho_\sigma) \sum_{i=0}^{m} C_{X,i} u^i \tag{A.1}$$

$$E_{C\sigma\sigma}^{GGA} = \int d\boldsymbol{r} \cdot e_{C\sigma\sigma}^{LSDA}(\rho_{\sigma\sigma}) \sum_{i=0}^{m} C_{C\sigma\sigma,i} u^i \tag{A.2}$$

$$E_{C\alpha\beta}^{GGA} = \int d\boldsymbol{r} \cdot e_{C\alpha\beta}^{LSDA}(\rho_{\alpha\beta}) \sum_{i=0}^{m} C_{C\alpha\beta,i} u^i \tag{A.3}$$

where e_X^{LSDA} is the LSDA exchange factor, $e_{C\sigma\sigma}^{LSDA}$ the same-spin correlation factor, $e_{C\alpha\beta}^{LSDA}$ the opposite-spin correlation factor. The corresponding polynomial expansion coefficients are indicated with a capital C. u^i is a function of the reduced density gradient (s) weighted by an attenuation factor (γ) and it has the following general form:

$$u = \frac{\gamma s^2}{1+\gamma s^2}. \tag{A.4}$$

In ωB97X-dDsC, as in ωB97X-D,[155] the polynomial expansion in Equations A.1, A.2 and A.3 is truncated at the fourth power (m=4). Building upon the GGA form, the exchange contribution of the range-separated hybrid functional is constructed as follows:

$$E_X = \sum_\sigma \int d\boldsymbol{r} \cdot e_{X\sigma}^{LSDA-SR(\omega)}(\rho_\sigma) \sum_{i=0}^{m} C_{X,i} u^i + C_X^{HF} \cdot E_X^{HF-SR(\omega)} + E_X^{HF-LR(\omega)} \tag{A.5}$$

where C_X^{HF} is the parameter determining the fraction of exact (Hartree-Fock) exchange at short range.

Adding the contribution from the dDsC dispersion correction[169–171] to Equations A.2, A.3 and A.5 yields the full exchange-correlation ωB97X-dDsC functional:

$$E_{XC}^{\omega B97X-dDsC} = E_X + E_{C\sigma\sigma}^{GGA} + E_{C\alpha\beta}^{GGA} + E_{dDsC} \tag{A.6}$$

Each component of Equation A.6 is tuned by at least two adjustable parameters. Table A.1 resumes the numerical value of all the optimized coefficients for ωB97X-dDsC, ωB97X+dDsC (where only the dispersion correction is reparametrized), ωB97X-D and ωB97X-D3. The latter functionals are included for comparison purposes.

Table A.1 – Adjustable parameters of the ωB97X-D, ωB97X-D3, ωB97X-dDsC and ωB97X+dDsC functionals

	-D	-D3	-dDsC	+dDsC
$\mathrm{C_x^{HF}}$	0.222036	0.195728	0.202143	0.157706
$\mathrm{C_x^{0}}$	0.777964	0.804272	0.797857	0.842294
$\mathrm{C_x^{1}}$	0.661160	0.698900	-0.100588	0.726479
$\mathrm{C_x^{2}}$	0.574541	0.508940	2.371856	1.04760
$\mathrm{C_x^{3}}$	-5.25671	-3.744903	-0.099302	-5.70635
$\mathrm{C_x^{4}}$	11.6386	10.060790	1.647653	1.32794
$\mathrm{C_{c,\sigma\sigma}^{0}}$	1.000000	1.000000	1.000000	1.000000
$\mathrm{C_{c,\sigma\sigma}^{1}}$	-6.90539	2.433266	-5.161406	-4.33879
$\mathrm{C_{c,\sigma\sigma}^{2}}$	31.3343	-15.446008	21.971045	18.2308
$\mathrm{C_{c,\sigma\sigma}^{3}}$	-51.0533	17.644390	-36.945577	-31.7430
$\mathrm{C_{c,\sigma\sigma}^{4}}$	26.4423	-8.879494	20.00011	17.2901
$\mathrm{C_{c,\alpha\beta}^{0}}$	1.000000	1.000000	1.000000	1.000000
$\mathrm{C_{c,\alpha\beta}^{1}}$	1.794130	-4.868902	5.479632	2.37031
$\mathrm{C_{c,\alpha\beta}^{2}}$	-12.047700	21.295726	-28.015938	-11.3995
$\mathrm{C_{c,\alpha\beta}^{3}}$	14.084700	-36.020866	19.533086	-31.7430
$\mathrm{C_{c,\alpha\beta}^{4}}$	-8.50809	19.177018	3.351842	17.2901
ω	0.2	0.25	0.275752	0.3
a	6	-	-	-
$\mathrm{s_{r,6}}$	-	1.281	-	-
$\mathrm{s_{r,8}}$	-	1.094	-	-
ATT0	-	-	59.57	23.30
BTT0	-	-	1.25	0.757

A.3 The CryOrel Set

A dataset containing all relevant structural and benchmark data will be made available upon publication in the Materials Cloud public repository.

Table A.2 – Abbreviations used in the CryOrel set

	Abbreviation	Definition
Type I		
	DITT:	(diindeno)-dithienothiophene
	ETTDM-TTF:	(ethylenethio)(thiodimethylene)-tetrathiafulvalene
	FPP-DTT:	(perfluorophenyl)(phenyl)-dithienothiophene
Type II		
	BBBT:	benzo-bis(benzothiophene)
	BDT:	bis(dithiophene)
	DBT-Sulfone:	dibenzothiophene-Sulfone
Type III		
	BTTT:	bis(thiophene)-thienothiophene
	DBT:	dibenzothiophene
	QTH:	quaterthiophene

A.4 Spin Densities

The difference between the α- and the β-spin densities (spin-density) is a readily available probe of the charge (de-)localization in real-space. In Figure A.2, we report the spin-densities of three dimers from the CryOrel9 dataset, the first two being representative of the stacked and tilted class and the third (ETTDM-TTF) showing how structural asymmetry helps restoring the correct attribution of radical cation character. The two functionals chosen represent opposite bounds of the error on the interaction energies of the stacked dimers. In the upper panels of Figure A.2, it can be seen how ωB97M-V converges to the incorrect solution, attributing the majority of the spin-density to the neutral monomer. In contrast, both functionals correctly describe the spin-densities of the tilted and the asymmetric dimer.

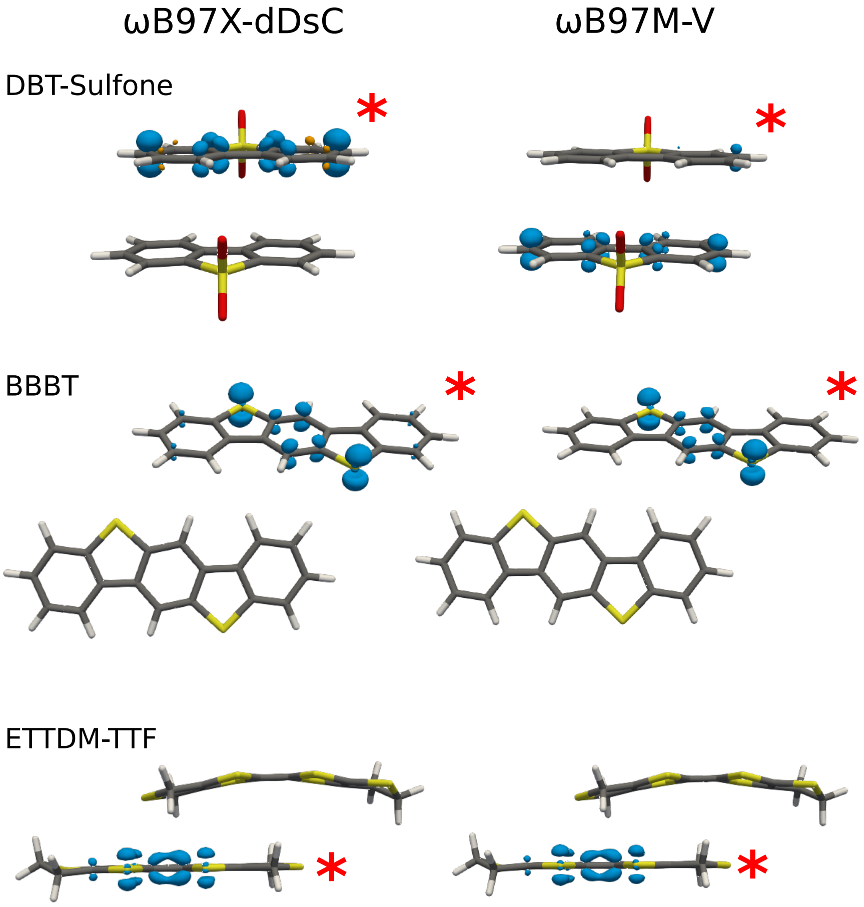


Figure A.2 – Spin-densities (isovalue: +0.01 e^{-}Bohr^{-3}) of DBT-Sulfone (stacked class), BBBT (tilted class) and ETTDM-TTF (stacked, but asymmetric) with two functionals (ωB97X-dDsC and ωB97M-V), representative of the lower and the upper bound of the interaction energy error. Red asterisks mark which monomer has been optimized as a radical cation.

B Energy Curvature: representations and local regression

B.1 Performance of different molecular representations in learning the average energy curvature

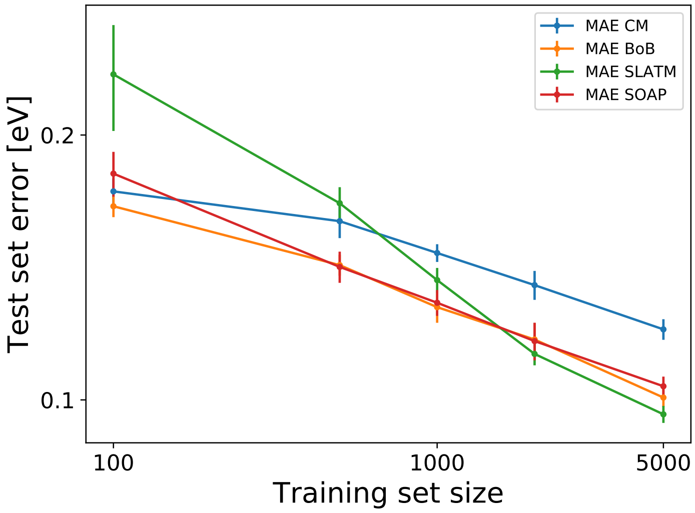


Figure B.1 – Learning curves of the average energy-curvature (C^N_{avg}) at LC-$omega$PBE/def2-SVP in function of the training set size. The learning exercise is reported for four different molecular representations: the Coulomb matrix (CM), the Bag of Bonds (BoB), the spectrum of London and Axilrod-Teller-Muto (SLATM) and the smooth overlap of atomic positions (SOAP). The error bars correspond to the standard deviation of the 10-fold cross-validation.

The performance of a machine-learning model targeting chemical properties depends strongly on the way the molecular information is represented.[59,250] A suitable representation constitutes in fact a meaningful relationship between the target property (herein the average energy curvature) and the molecular structure and composition. Over the last few years, several physically motivated molecular representations have been proposed, each of them including an increasing amount of chemical information.[25,27,38,59,244–251] Figure B.1 shows the performance in terms of mean absolute error of the average energy curvature computed at LC-ωPBE level using four different molecular representations: the Coulomb matrix (CM),[25] the Bag of Bonds (BoB),[247] the spectrum of London and Axilrod-Teller-Muto (SLATM)[250,251] and the smooth overlap of atomic positions (SOAP).[59]

Overall, the SLATM representation leads to the lowest mean absolute error at the full training set and to the steepest learning curve. The final accuracy of the other representations tested is nevertheless comparable, resulting in particularly small deviations ranging from 4 meV (BoB) to 24 meV (CM).

B.2 Local framework for the regression of the energy curvature

The average energy curvature (C^N_{avg}) for a fixed functional is a molecular property, whose partitioning into atomic contributions cannot be defined uniquely. Instead of imposing *a priori* a decomposition scheme, we construct a machine-learning model able to perform the regression and simultaneously find the most suitable atomic partitioning of C^N_{avg}. Figure B.2 is a schematic representation of such a regression framework. First, the molecular information is vectorized as a collection of atomic environments using the aSLATM representation. Then, Gaussian similarity kernels are evaluated between all the local environments, resulting in a N_{at} X N_{at} matrix, where N_{at} is the number of atoms in the training set. Since the dimensionality of the target C^N_{avg} is instead equal to the number of compounds, the lines of the kernel matrix are averaged for the atoms belonging to each molecule. Building a molecular similarity measure by averaging its local atomic contributions is not a novelty, but it represents the most straightforwards solution when evaluating the similarity of different compounds on the basis of a local representation.[599]

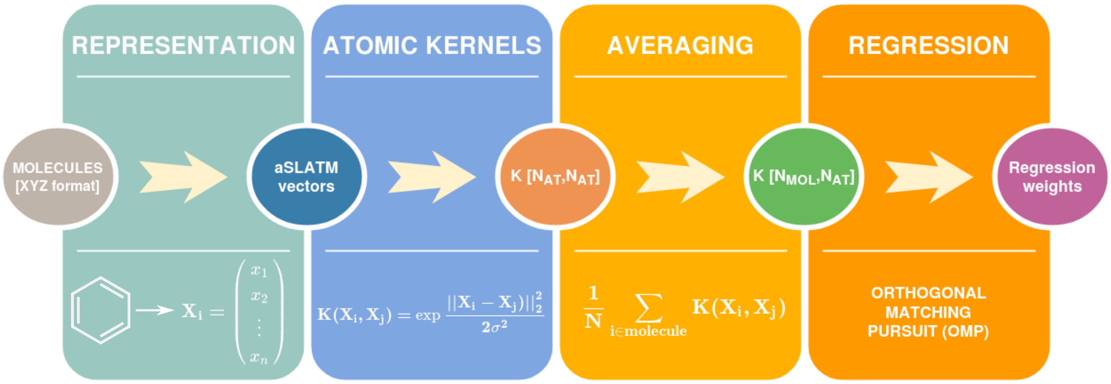

REPRESENTATION
ATOMIC KERNELS
AVERAGING
REGRESSION
MOLECULES [XYZ format]
aSLATM vectors
K [N_{AT},N_{AT}]
K [N_{MOL},N_{AT}]
Regression weights
$X_i = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$
$K(X_i, X_j) = \exp \frac{||X_i - X_j||_2^2}{2\sigma^2}$
$\frac{1}{N} \sum_{i \in \text{molecule}} K(X_i, X_j)$
ORTHOGONAL MATCHING PURSUIT (OMP)

Figure B.2 – Schematic representation of the local framework for the regression of the energy curvature.

The rectangular kernel resulting from the averaging procedure cannot be directly inverted to solve the regression problem. This over-complete (redundant) problem can be tackled using a sparse regression technique originally developed for signal recovery: the orthogonal matching pursuit (OMP) algorithm.[600,601] Given a fixed number of non-zero parameters($n_{NonZeroCoeff}$), this method is able to approximate the optimum regression weights vector (ω_{sol}) by

$$\omega_{sol} = arg\ min\ ||Y - K\omega||_2^2 \quad \text{Subject to} \quad ||\omega||_0 \leq n_{NonZeroCoeff} \tag{B.1}$$

where K is a over-complete kernel and Y is the regression target. For the model presented in this work the 300 non-zero coefficients were found to be optimal.

Bibliography

[1] Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry - Introduction to Advanced Electronic Structure Theory*; Mc Graw-Hill Publishing Company: New York, 1982.

[2] Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; John Wiley & Sons, Ltd: Chichester, UK, 2000.

[3] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York: New York, NY, 2009.

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

[5] Kohn, W.; Becke, A. D.; Parr, R. G. Density Functional Theory of Electronic Structure. *J. Phys. Chem.* **1996**, *100*, 12974–12980.

[6] Burke, K. Perspective on Density Functional Theory. *J. Chem. Phys.* **2012**, *136*, 150901.

[7] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289–320.

[8] Becke, A. D. Perspective: Fifty Years of Density-Functional Theory in Chemical Physics. *J. Chem. Phys.* **2014**, *140*, 18A301.

[9] Jones, R. O. Density Functional Theory: Its Origins, Rise to Prominence, and Future. *Rev. Mod. Phys.* **2015**, *87*, 897–923.

[10] Yu, H. S.; Li, S. L.; Truhlar, D. G. Perspective: Kohn-Sham Density Functional Theory Descending a Staircase. *J. Chem. Phys.* **2016**, *145*, 130901.

[11] Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

[12] Perdew, J. P.; Zunger, A. Self-Interaction Correction to Density-Functional Approximations for Many-Electron Systems. *Phys. Rev. B* **1981**, *23*, 5048–5079.

[13] Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. Density Functional Theory for the Fractional Particle Number: Derivative Discontinuities of the Energy. *Phys. Rev. Lett.* **1982**,

[14] Zhang, Y.; Yang, W. A Challenge for Density Functionals: Self-Interaction Error Increases for Systems with a Noninteger Number of Electrons. *J. Chem. Phys.* **1998**, *109*, 2604–2608.

[15] Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Many-Electron Self-Interaction Error in Approximate Density Functionals. *J. Chem. Phys.* **2006**, *125*, 201102.

[16] Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. Spurious Fractional Charge on Dissociated Atoms: Pervasive and Resilient Self-Interaction Error of Common Density Functionals. *J. Chem. Phys.* **2006**, *125*, 194112.

[17] Cohen, A. J.; Mori-Sanchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792–794.

[18] Kristyán, S.; Pulay, P. Can (semi)local Density Functional Theory Account for the London Dispersion Forces? *Chem. Phys. Lett.* **1994**, *229*, 175–180.

[19] Pérez-Jordá, J. M.; Becke, A. A Density-Functional Study of Van Der Waals Forces: Rare Gas Diatomics. *Chem. Phys. Lett.* **1995**, *233*, 134–137.

[20] Hobza, P.; Sponer, J.; Reschel, T. Density Functional Theory and Molecular Clusters. *J. Comput. Chem.* **1995**, *16*, 1315–1325.

[21] Meijer, E. J.; Sprik, M. A Density-functional Study of the Intermolecular Interactions of Benzene. *J. Chem. Phys.* **1996**, *105*, 8684–8689.

[22] Perdew, J. P.; Savin, A.; Burke, K. Escaping the Symmetry Dilemma Through a Pair-Density Interpretation of Spin-Density Functional Theory. *Phys. Rev. A* **1995**, *51*, 4531–4541.

[23] Cremer, D. Density Functional Theory: Coverage of Dynamic and Non-Dynamic Electron Correlation Effects. *Mol. Phys.* **2001**, *99*, 1899–1940.

[24] Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.

[25] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Lilienfeld, V.; Anatole, O. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 58301.

[26] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.

[27] Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals. *Phys. Rev. Lett.* **2016**, *117*, 135502.

[28] Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower Than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

[29] Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9*, 7069–7077.

[30] von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. *Preprint* **2019**, *arXiv:1911.10084*.

[31] Liang, J.; Xu, Y.; Liu, R.; Zhu, X. QM-Sym, a Symmetrized Quantum Chemistry Database of 135 Kilo Molecules. *Sci. Data* **2019**, *6*, 213.

[32] Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8*, 872.

[33] Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. *ACS Centr. Sci.* **2019**, *5*, 57–64.

[34] Fabrizio, A.; Grisafi, A.; Meyer, B.; Ceriotti, M.; Corminboeuf, C. Electron Density Learning of Non-Covalent Systems. *Chem. Sci.* **2019**, *10*, 9424–9432.

[35] 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*, 5024.

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

[37] Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-The-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.

[38] Botu, V.; Ramprasad, R. Adaptive Machine Learning Framework to Accelerate Ab Initio Molecular Dynamics. *Int. J. Quantum Chem.* **2015**, *115*, 1074–1083.

[39] Krems, R. V. Bayesian Machine Learning for Quantum Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13392–13410.

[40] Westermayr, J.; Gastegger, M.; Menger, M. F. S. J.; Mai, S.; González, L.; Marquetand, P. Machine Learning Enables Long Time Scale Molecular Photodynamics Simulations. *Chem. Sci.* **2019**, *10*, 8100–8107.

[41] von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angew. Chemie Int. Ed.* **2018**, *57*, 4164–4169.

[42] Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL Revisited: Faster and More Accurate Quantum Machine Learning. *J. Chem. Phys.* **2020**, *152*, 044107.

[43] Rupp, M.; von Lilienfeld, O. A.; Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *J. Chem. Phys.* **2018**, *148*, 241401.

[44] Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3*, e1603015.

[45] Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields. *Nat. Comm.* **2018**, *9*, 3887.

[46] Bereau, T.; Andrienko, D.; von Lilienfeld, O. A. Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules. *J. Chem. Theory Comput.* **2015**, *11*, 3225–3233.

[47] Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, 036002.

[48] Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate Molecular Polarizabilities with Coupled Cluster Theory and Machine Learning. *Proc. Natl. Acad. Sci.* **2019**, *116*, 3401–3406.

[49] Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

[50] Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, Without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

[51] Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chemie Int. Ed.* **2017**, *56*, 12828–12840.

[52] Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018**, *148*, 241733.

[53] Schmidt, E.; Fowler, A. T.; Elliott, J. A.; Bristowe, P. D. Learning Models for Electron Densities with Bayesian Regression. *Comput. Mater. Sci.* **2018**, *149*, 250–258.

[54] Alred, J. M.; Bets, K. V.; Xie, Y.; Yakobson, B. I. Machine Learning Electron Density in Sulfur Crosslinked Carbon Nanotubes. *Compos. Sci. Technol.* **2018**, *166*, 3–9.

[55] Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the Electronic Structure Problem with Machine Learning. *npj Comput. Mater.* **2019**, *5*, 22.

[56] Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.

[57] Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.

[58] Hollingsworth, J.; Baker, T. E.; Burke, K. Can Exact Conditions Improve Machine-Learned Density Functionals? *J. Chem. Phys.* **2018**, *148*, 241743.

[59] Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 184115.

[60] Steinmann, S. N.; Corminboeuf, C. Exploring the Limits of Density Functional Approximations for Interaction Energies of Molecular Precursors to Organic Electronics. *J. Chem. Theory Comput.* **2012**, *8*, 4305–16.

[61] Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G.; Vanommeslaeghe, K.; MacKerell, A. D.; Merz, K. M.; Sherrill, C. D. The BioFragment Database (BFDb): An Open-Data Platform for Computational Chemistry Analysis of Noncovalent Interactions. *J. Chem. Phys.* **2017**, *147*, 161727.

[62] Stein, T.; Autschbach, J.; Govind, N.; Kronik, L.; Baer, R. Curvature and Frontier Orbital Energies in Density Functional Theory. *J. Phys. Chem. Lett.* **2012**, *3*, 3740–3744.

[63] Kraisler, E.; Kronik, L. Piecewise Linearity of Approximate Density Functionals Revisited: Implications for Frontier Orbital Energies. *Phys. Rev. Lett.* **2013**, *110*, 126403.

[64] Hait, D.; Head-Gordon, M. Delocalization Errors in Density Functional Theory Are Essentially Quadratic in Fractional Occupation Number. *J. Phys. Chem. Lett.* **2018**, *9*, 6280–6288.

[65] Perdew, J. P.; Levy, M. Comment on "Significance of the Highest Occupied Kohn-Sham Eigenvalue". *Phys. Rev. B* **1997**, *56*, 16021–16028.

[66] Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. Density Functionals That Are One- and Two- Are Not Always Many-Electron Self-Interaction-Free, As Shown for H2+, He2+, LiH+, and Ne2+. *J. Chem. Phys.* **2007**, *126*, 104102.

[67] Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P. Tests of Functionals for Systems with Fractional Electron Number. *J. Chem. Phys.* **2007**, *126*, 154109.

[68] Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Localization and Delocalization Errors in Density Functional Theory and Implications for Band-Gap Prediction. *Phys. Rev. Lett.* **2008**, *100*, 146401.

[69] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Fractional Charge Perspective on the Band Gap in Density-Functional Theory. *Phys. Rev. B* **2008**, *77*, 115123.

[70] Haunschild, R.; Henderson, T. M.; Jiménez-Hoyos, C. A.; Scuseria, G. E. Many-Electron Self-Interaction and Spin Polarization Errors in Local Hybrid Density Functionals. *J. Chem. Phys.* **2010**, *133*, 134116.

[71] Ayers, P. W.; Levy, M. Sum Rules for Exchange and Correlation Potentials. *J. Chem. Phys.* **2001**, *115*, 4438–4443.

[72] Perdew, J. P. *Density Funct. Methods Phys.*; Springer US, 1985; pp 265–308.

[73] Johnson, E. R.; Otero-de-la Roza, A.; Dale, S. G. Extreme Density-Driven Delocalization Error for a Model Solvated-Electron System. *J. Chem. Phys.* **2013**, *139*, 184116.

[74] Johnson, E. R.; Salamone, M.; Bietti, M.; DiLabio, G. A. Modeling Noncovalent Radical–Molecule Interactions Using Conventional Density-Functional Theory: Beware Erroneous Charge Transfer. *J. Phys. Chem. A* **2013**, *117*, 947–952.

[75] van Gisbergen, S. J. A.; Schipper, P. R. T.; Gritsenko, O. V.; Baerends, E. J.; Snijders, J. G.; Champagne, B.; Kirtman, B. Electric Field Dependence of the Exchange-Correlation Potential in Molecular Chains. *Phys. Rev. Lett.* **1999**, *83*, 694–697.

[76] Mori-Sánchez, P.; Wu, Q.; Yang, W. Accurate Polymer Polarizabilities with Exact Exchange Density-Functional Theory. *J. Chem. Phys.* **2003**, *119*, 11001–11004.

[77] Perdew, J. P.; Levy, M. Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities. *Phys. Rev. Lett.* **1983**, *51*, 1884–1887.

[78] Allen, M. J.; Tozer, D. J. Eigenvalues, Integer Discontinuities and NMR Shielding Constants in Kohn—Sham Theory. *Mol. Phys.* **2002**, *100*, 433–439.

[79] Kümmel, S.; Kronik, L. Orbital-Dependent Density Functionals: Theory and Applications. *Rev. Mod. Phys.* **2008**, *80*, 3–60.

[80] Görling, A. Exchange-Correlation Potentials with Proper Discontinuities for Physically Meaningful Kohn-Sham Eigenvalues and Band Structures. *Phys. Rev. B* **2015**, *91*, 245120.

[81] Becke, A. D. A New Mixing of Hartree–Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98*, 1372.

[82] Becke, A. D. Density-Functional Thermochemistry. III. the Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648.

[83] Johnson, E. R.; Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Delocalization Errors in Density Functionals and Implications for Main-Group Thermochemistry. *J. Chem. Phys.* **2008**, *129*, 204112.

[84] Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A. Global Hybrid Functionals: A Look at the Engine Under the Hood. *J. Chem. Theory Comput.* **2010**, *6*, 3688–3703.

[85] Sun, H.; Autschbach, J. Influence of the Delocalization Error and Applicability of Optimal Functional Tuning in Density Functional Calculations of Nonlinear Optical Properties of Organic Donor-Acceptor Chromophores. *ChemPhysChem* **2013**, *14*, 2450–2461.

[86] Autschbach, J.; Srebro, M. Delocalization Error and “Functional Tuning” in Kohn–Sham Calculations of Molecular Properties. *Acc. Chem. Res.* **2014**, *47*, 2592–2602.

[87] Harris, J.; Jones, R. O. The Surface Energy of a Bounded Electron Gas. *J. Phys. F: Met. Phys.* **1974**, *4*, 1170–1186.

[88] Gunnarsson, O.; Lundqvist, B. I. Exchange and Correlation in Atoms, Molecules, and Solids by the Spin-Density-Functional Formalism. *Phys. Rev. B* **1976**, *13*, 4274–4298.

[89] Langreth, D. C.; Perdew, J. P. Exchange-Correlation Energy of a Metallic Surface: Wave-Vector Analysis. *Phys. Rev. B* **1977**, *15*, 2884–2901.

[90] Harris, J. Adiabatic-Connection Approach to Kohn-Sham Theory. *Phys. Rev. A* **1984**, *29*, 1648–1659.

[91] Levy, M.; Perdew, J. P.; Sahni, V. Exact Differential Equation for the Density and Ionization Energy of a Many-Particle System. *Phys. Rev. A* **1984**, *30*, 2745–2748.

[92] Almbladh, C.-O.; von Barth, U. Exact Results for the Charge and Spin Densities, Exchange-Correlation Potentials, and Density-Functional Eigenvalues. *Phys. Rev. B* **1985**, *31*, 3231–3244.

[93] van Leeuwen, R.; Baerends, E. J. Exchange-Correlation Potential with Correct Asymptotic Behavior. *Phys. Rev. A* **1994**, *49*, 2421–2431.

[94] Burke, K.; Perdew, J. P.; Langreth, D. C. Is the Local Density Approximation Exact for Short Wavelength Fluctuations? *Phys. Rev. Lett.* **1994**, *73*, 1283–1286.

[95] Burke, K.; Perdew, J. P. Real-Space Analysis of the Exchange-Correlation Energy. *Int. J. Quantum Chem.* **1995**, *56*, 199–210.

[96] Toulouse, J.; Colonna, F.; Savin, A. Long-Range–short-Range Separation of the Electron-Electron Interaction in Density-Functional Theory. *Phys. Rev. A* **2004**, *70*, 062505.

[97] Kutzelnigg, W.; Morgan, J. D. Erratum: Rates of Convergence of the Partial-wave Expansions of Atomic Correlation Energies [J. Chem. Phys. 96 , 4484 (1992)]. *J. Chem. Phys.* **1992**, *97*, 8821–8821.

[98] Kutzelnigg, W.; Morgan, J. D. Rates of Convergence of the Partial-wave Expansions of Atomic Correlation Energies. *J. Chem. Phys.* **1992**, *96*, 4484–4508.

[99] Savin, A.; Flad, H.-J. Density Functionals for the Yukawa Electron-Electron Interaction. *Int. J. Quantum Chem.* **1995**, *56*, 327–332.

[100] Savin, A. *Theor. Comput. Chem.*; 1996; pp 327–357.

[101] Gill, P. M.; Adamson, R. D. A Family of Attenuated Coulomb Operators. *Chem. Phys. Lett.* **1996**, *261*, 105–110.

[102] Livshits, E.; Baer, R. A Well-Tempered Density Functional Theory of Electrons in Molecules. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932.

[103] Livshits, E.; Baer, R. A Density Functional Theory for Symmetric Radical Cations from Bonding to Dissociation. *J. Phys. Chem. A* **2008**, *112*, 12789–12791.

[104] Baer, R.; Livshits, E.; Salzner, U. Tuned Range-Separated Hybrids in Density Functional Theory. *Annu. Rev. Phys. Chem.* **2010**, *61*, 85–109.

[105] Gledhill, J. D.; Peach, M. J. G.; Tozer, D. J. Assessment of Tuning Methods for Enforcing Approximate Energy Linearity in Range-Separated Hybrid Functionals. *J. Chem. Theory Comput.* **2013**, *9*, 4414–4420.

[106] Karolewski, A.; Kronik, L.; Kümmel, S. Using Optimally Tuned Range Separated Hybrid Functionals in Ground-State Calculations: Consequences and Caveats. *J. Chem. Phys.* **2013**, *138*, 204115.

[107] Henderson, T. M.; Janesko, B. G.; Scuseria, G. E. Range Separation and Local Hybridization in Density Functional Theory. *J. Phys. Chem. A* **2008**,

[108] Jaramillo, J.; Scuseria, G. E.; Ernzerhof, M. Local Hybrid Functionals. *J. Chem. Phys.* **2003**, *118*, 1068–1073.

[109] Maier, T. M.; Arbuznikov, A. V.; Kaupp, M. Local Hybrid Functionals: Theory, Implementation, and Performance of an Emerging New Tool in Quantum Chemistry and Beyond. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2019**, *9*, e1378.

[110] Krukau, A. V.; Scuseria, G. E.; Perdew, J. P.; Savin, A. Hybrid Functionals with Local Range Separation. *J. Chem. Phys.* **2008**, *129*, 124103.

[111] Janesko, B. G.; Krukau, A. V.; Scuseria, G. E. Self-Consistent Generalized Kohn-Sham Local Hybrid Functionals of Screened Exchange: Combining Local and Range-Separated Hybridization. *J. Chem. Phys.* **2008**, *129*, 124110.

[112] Haunschild, R.; Scuseria, G. E. Range-Separated Local Hybrids. *J. Chem. Phys.* **2010**, *132*, 224106.

[113] Polo, V.; Gräfenstein, J.; Kraka, E.; Cremer, D. Influence of the Self-Interaction Error on the Structure of the DFT Exchange Hole. *Chem. Phys. Lett.* **2002**, *352*, 469–478.

[114] Polo, V.; Kraka, E.; Cremer, D. Electron Correlation and the Self-Interaction Error of Density Functional Theory. *Mol. Phys.* **2002**, *100*, 1771–1790.

[115] Ciofini, I.; Adamo, C.; Chermette, H. Self-Interaction Error in Density Functional Theory: A Mean-Field Correction for Molecules and Large Systems. *Chem. Phys.* **2005**, *309*, 67–76.

[116] Lundberg, M.; Siegbahn, P. E. M. Quantifying the Effects of the Self-Interaction Error in DFT: When Do the Delocalized States Appear? *J. Chem. Phys.* **2005**, *122*, 224103.

[117] Gräfenstein, J.; Cremer, D. The Self-Interaction Error and the Description of Non-Dynamic Electron Correlation in Density Functional Theory. *Theor. Chem. Acc.* **2009**, *123*, 171–182.

[118] Bao, J. L.; Gagliardi, L.; Truhlar, D. G. Self-Interaction Error in Density Functional Theory: An Appraisal. *J. Phys. Chem. Lett.* **2018**, *9*, 2353–2358.

[119] Fermi, E.; Amaldi, E. Le Orbite ∞-S Degli Elementi. *Mem. Accad. d'Italia* **1934**, *6*, 119.

[120] Slater, J. C.; Wood, J. H. Statistical Exchange and the Total Energy of a Crystal. *Int. J. Quantum Chem.* **1971**, *4*, 3–34.

[121] Perdew, J. Orbital Functional for Exchange and Correlation: Self-Interaction Correction to the Local Density Approximation. *Chem. Phys. Lett.* **1979**, *64*, 127–130.

[122] Perdew, J. P. *Adv. Quantum Chem.*; 1990; pp 113–134.

[123] Tsuneda, T.; Hirao, K. Self-Interaction Corrections in Density Functional Theory. *J. Chem. Phys.* **2014**, *140*, 18A513.

[124] Heaton, R. A.; Harrison, J. G.; Lin, C. C. Self-Interaction Correction for Density-Functional Theory of Electronic Energy Bands of Solids. *Phys. Rev. B* **1983**, *28*, 5992–6007.

[125] Johnson, B. G.; Gonzales, C. A.; Gill, P. M. W.; Pople, J. A. A Density Functional Study of the Simplest Hydrogen Abstraction Reaction. Effect of Self-Interaction Correction. *Chem. Phys. Lett.* **1994**, *221*, 100–108.

[126] Pederson, M. R.; Lin, C. C. Localized and Canonical Atomic Orbitals in Self-interaction Corrected Local Density Functional Approximation. *J. Chem. Phys.* **1988**, *88*, 1807–1817.

[127] Tong, X.-M.; Chu, S.-I. Density-Functional Theory with Optimized Effective Potential and Self-Interaction Correction for Ground States and Autoionizing Resonances. *Phys. Rev. A* **1997**, *55*, 3406–3416.

[128] Pederson, M. R.; Ruzsinszky, A.; Perdew, J. P. Communication: Self-Interaction Correction with Unitary Invariance in Density Functional Theory. *J. Chem. Phys.* **2014**, *140*, 121103.

[129] Lehtola, S.; Jónsson, H. Variational, Self-Consistent Implementation of the Perdew–Zunger Self-Interaction Correction with Complex Optimal Orbitals. *J. Chem. Theory Comput.* **2014**, *10*, 5324–5337.

[130] Lehtola, S.; Head-Gordon, M.; Jónsson, H. Complex Orbitals, Multiple Local Minima, and Symmetry Breaking in Perdew–Zunger Self-Interaction Corrected Density Functional Theory Calculations. *J. Chem. Theory Comput.* **2016**, *12*, 3195–3207.

[131] Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Self-Interaction-Free Exchange-Correlation Functional for Thermochemistry and Kinetics. *J. Chem. Phys.* **2006**, *124*, 091102.

[132] Zhao, Y.; Truhlar, D. G. Density Functional for Spectroscopy: No Long-Range Self-Interaction Error, Good Performance for Rydberg and Charge-Transfer States, and Better Performance on Average Than B3LYP for Ground States. *J. Phys. Chem. A* **2006**, *110*, 13126–13130.

[133] Grimme, S. Density Functional Theory with London Dispersion Corrections. *WIREs Comput. Mol. Sci.* **2011**, *1*, 211–228.

[134] Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chem. Rev.* **2016**, *116*, 5105–5154.

[135] Dion, M.; Rydberg, H.; Schroeder, E.; Langreth, D. C.; Lundqvist, B. I. Van Der Waals Density Functional for General Geometries. *Phys. Rev. Lett.* **2004**, *92*, 246401–1.

[136] Lee, K.; Murray, É. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. Higher-Accuracy Van Der Waals Density Functional. *Phys. Rev. B* **2010**, *82*, 081101.

[137] Vydrov, O. A.; Van Voorhis, T. Nonlocal Van Der Waals Density Functional: The Simpler the Better. *J. Chem. Phys.* **2010**, *133*, 244103.

[138] von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. Optimization of Effective Atom Centered Potentials for London Dispersion Forces in Density Functional Theory. *Phys. Rev. Lett.* **2004**, *93*, 153004.

[139] DiLabio, G. A. Accurate Treatment of Van Der Waals Interactions Using Standard Density Functional Theory Methods with Effective Core-Type Potentials: Application to Carbon-Containing Dimers. *Chem. Phys. Lett.* **2008**, *455*, 348–353.

[140] Zhao, Y.; Truhlar, D. G. Density Functionals with Broad Applicability in Chemistry. *Acc. Chem. Res.* **2008**, *41*, 157–167.

[141] Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. Dispersion Interactions in Density-Functional Theory. *J. Phys. Org. Chem.* **2009**, *22*, 1127–1135.

[142] Cooper, V.; Kong, L.; Langreth, D. Computing Dispersion Interactions in Density Functional Theory. *Phys. Procedia* **2010**, *3*, 1417–1430.

[143] Hirata, S.; He, X.; Hermes, M. R.; Willow, S. Y. Second-Order Many-Body Perturbation Theory: An Eternal Frontier. *J. Phys. Chem. A* **2014**, *118*, 655–672.

[144] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

[145] Tkatchenko, A.; Distasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body Van Der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.

[146] Otero-de-la Roza, A.; Johnson, E. R. Many-Body Dispersion Interactions from the Exchange-Hole Dipole Moment Model. *J. Chem. Phys.* **2013**, *138*, 054103.

[147] Koide, A. A New Expansion for Dispersion Forces and Its Application. *J. Phys. B At. Mol. Phys.* **1976**, *9*, 3173–3183.

[148] Wu, Q.; Yang, W. Empirical Correction to Density Functional Theory for Van Der Waals Interactions. *J. Chem. Phys.* **2002**, *116*, 515–524.

[149] Spackman, M. A. Time-dependent Hartree–Fock Second-order Molecular Properties with a Moderately Sized Basis Set. II. Dispersion Coefficients. *J. Chem. Phys.* **1991**, *94*, 1295–1305.

[150] Stanton, J. F. Calculation of C6 Dispersion Constants with Coupled-Cluster Theory. *Phys. Rev. A* **1994**, *49*, 1698–1703.

[151] Halgren, T. A. The Representation of Van Der Waals (vdW) Interactions in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and VdW Parameters. *J. Am. Chem. Soc.* **1992**, *114*, 7827–7843.

[152] Tang, K. T.; Toennies, J. P. An Improved Simple Model for the Van Der Waals Potential Based on Universal Damping Functions for the Dispersion Coefficients. *J. Chem. Phys.* **1984**, *80*, 3726.

[153] Tang, K. T.; Toennies, J. P.; Yiu, C. L. Accurate Analytical He-He Van Der Waals Potential Based on Perturbation Theory. *Phys. Rev. Lett.* **1995**, *74*, 1546–1549.

[154] Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

[155] Chai, J.-D.; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom-Atom Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

[156] Liu, Y.; Goddard, W. A. I. A Universal Damping Function for Empirical Dispersion Correction on Density Functional Theory. *Mater. Trans.* **2009**, *50*, 1664–1670.

[157] Becke, A. D.; Johnson, E. R. A Density-Functional Model of the Dispersion Interaction. *J. Chem. Phys.* **2005**, *123*, 154101.

[158] Johnson, E. R.; Becke, A. D. A Post-Hartree–Fock Model of Intermolecular Interactions. *J. Chem. Phys.* **2005**, *123*, 024101.

[159] Johnson, E. R.; Becke, A. D. A Post-Hartree-Fock Model of Intermolecular Interactions: Inclusion of Higher-Order Corrections. *J. Chem. Phys.* **2006**, *124*, 174104.

[160] Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

[161] Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A Generally Applicable Atomic-Charge Dependent London Dispersion Correction. *J. Chem. Phys.* **2019**, *150*, 154122.

[162] Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.

[163] Silvestrelli, P. L. van Der Waals Interactions in Density Functional Theory Using Wannier Functions. *J. Phys. Chem. A* **2009**, *113*, 5224–5234.

[164] Ambrosetti, A.; Silvestrelli, P. L. van Der Waals Interactions in Density Functional Theory Using Wannier Functions: Improved C6 and C3 Coefficients by a Different Approach. *Phys. Rev. B* **2012**, *85*, 073101.

[165] Silvestrelli, P. L. Van Der Waals Interactions in Density Functional Theory by Combining the Quantum Harmonic Oscillator-Model with Localized Wannier Functions. *J. Chem. Phys.* **2013**, *139*, 054106.

[166] Becke, A. D.; Johnson, E. R. Exchange-Hole Dipole Moment and the Dispersion Interaction. *J. Chem. Phys.* **2005**, *127*, 154108.

[167] Becke, A. D.; Johnson, E. R. Exchange-Hole Dipole Moment and the Dispersion Interaction: High-Order Dispersion Coefficients. *J. Chem. Phys.* **2006**, *124*, 014104.

[168] Becke, A. D.; Johnson, E. R. Exchange-Hole Dipole Moment and the Dispersion Interaction Revisited. *The Journal of Chemical Physics* **2007**, *127*, 154108.

[169] Steinmann, S. N.; Corminboeuf, C. A System-Dependent Density-Based Dispersion Correction. *J. Chem. Theory Comput.* **2010**, *6*, 1990–2001.

[170] Steinmann, S. N.; Corminboeuf, C. A Generalized-Gradient Approximation Exchange Hole Model for Dispersion Coefficients. *J. Chem. Phys.* **2011**, *134*, 44117.

[171] Steinmann, S. N.; Corminboeuf, C. Comprehensive Benchmarking of a Density-Dependent Dispersion Correction. *J. Chem. Theory Comput.* **2011**, *7*, 3567–3577.

[172] Sato, T.; Nakai, H. Density Functional Method Including Weak Interactions: Dispersion Coefficients Based on the Local Response Approximation. *J. Chem. Phys.* **2009**, *131*, 224104.

[173] Sato, T.; Nakai, H. Local Response Dispersion Method. II. Generalized Multicenter Interactions. *J. Chem. Phys.* **2010**, *133*, 194101.

[174] Becke, A. D.; Roussel, M. R. Exchange Holes in Inhomogeneous Systems: A Coordinate-Space Model. *Phys. Rev. A* **1989**, *39*, 3761–3767.

[175] Ángyán, J. G. on the Exchange-Hole Model of London Dispersion Forces. *J. Chem. Phys.* **2007**, *127*, 024108.

[176] Slater, J. C.; Kirkwood, J. G. The Van Der Waals Forces in Gases. *Phys. Rev.* **1931**, *37*, 682–697.

[177] Cambi, R.; Cappelletti, D.; Liuti, G.; Pirani, F. Generalized Correlations in Terms of Polarizability for Van Der Waals Interaction Potential Parameter Calculations. *J. Chem. Phys.* **1991**, *95*, 1852–1861.

[178] Brinck, T.; Murray, J. S.; Politzer, P. Polarizability and Volume. *J. Chem. Phys.* **1993**, *98*, 4305–4306.

[179] Levy, M. Universal Variational Functionals of Electron Densities, First-Order Density Matrices, and Natural Spin-Orbitals and Solution of the v-Representability Problem. *Proc. Natl. Acad. Sci.* **1979**, *76*, 6062–6065.

[180] Pulay, P. Convergence Acceleration of Iterative Sequences. the Case of Scf Iteration. *Chem. Phys. Lett.* **1980**, *73*, 393–398.

[181] Bacskay, G. B. A Quadratically Convergent Hartree—Fock (QC-SCF) Method. Application to Closed Shell Systems. *Chem. Phys.* **1981**, *61*, 385–404.

[182] Bacskay, G. B. A Quadritically Convergent Hartree-Fock (QC-SCF) Method. Application to Open Shell Orbital Optimization and Coupled Perturbed Hartree-Fock Calculations. *Chem. Phys.* **1982**, *65*, 383–396.

[183] Pulay, P. ImprovedSCF Convergence Acceleration. *J. Comput. Chem.* **1982**, *3*, 556–560.

[184] Chaban, G.; Schmidt, M. W.; Gordon, M. S. Approximate Second Order Method for Orbital Optimization of SCF and MCSCF Wavefunctions. *Theor. Chem. Accounts Theory, Comput. Model. (Theoretica Chim. Acta)* **1997**, *97*, 88–95.

[185] Neese, F. Approximate Second-Order SCF Convergence for Spin Unrestricted Wavefunctions. *Chem. Phys. Lett.* **2000**, *325*, 93–98.

[186] Kudin, K. N.; Scuseria, G. E.; Cancès, E. A Black-Box Self-Consistent Field Convergence Algorithm: One Step Closer. *J. Chem. Phys.* **2002**, *116*, 8255.

[187] Hu, X.; Yang, W. Accelerating Self-Consistent Field Convergence with the Augmented Roothaan–Hall Energy Function. *J. Chem. Phys.* **2010**, *132*, 054109.

[188] Chen, Y. K.; Wang, Y. A. LISTb: A Better Direct Approach to LIST. *J. Chem. Theory Comput.* **2011**, *7*, 3045–3048.

[189] Wang, Y. A.; Yam, C. Y.; Chen, Y. K.; Chen, G. Communication: Linear-Expansion Shooting Techniques for Accelerating Self-Consistent Field Convergence. *J. Chem. Phys.* **2011**, *134*, 241103.

[190] Wolfsberg, M.; Helmholz, L. The Spectra and Electronic Structure of the Tetrahedral Ions MnO_4^-, CrO_4^{--}, and ClO_4^-. *J. Chem. Phys.* **1952**, *20*, 837–843.

[191] Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39*, 1397–1412.

[192] Almlöf, J.; Faegri, K.; Korsell, K. Principles for a Direct SCFapproach to LCAO-MO Ab-Initio Calculations. *J. Comput. Chem.* **1982**, *3*, 385–399.

[193] Van Lenthe, J. H.; Zwaans, R.; Van Dam, H. J. J.; Guest, M. F. Starting SCF Calculations by Superposition of Atomic Densities. *J. Comput. Chem.* **2006**, *27*, 926–932.

[194] Lehtola, S. Assessment of Initial Guesses for Self-Consistent Field Calculations. Superposition of Atomic Potentials: Simple Yet Efficient. *J. Chem. Theory Comput.* **2019**, *15*, 1593–1604.

[195] Becke, A. D. A Multicenter Numerical Integration Scheme for Polyatomic Molecules. *J. Chem. Phys.* **1988**, *88*, 2547–2553.

[196] Murray, C. W.; Handy, N. C.; Laming, G. J. Quadrature Schemes for Integrals of Density Functional Theory. *Mol. Phys.* **1993**, *78*, 997–1014.

[197] Gill, P. M. W.; Chien, S.-H. Radial Quadrature for Multiexponential Integrands. *J. Comput. Chem.* **2003**, *24*, 732–740.

[198] Köster, A. M.; Flores-Moreno, R.; Reveles, J. U. Efficient and Reliable Numerical Integration of Exchange-Correlation Energies and Potentials. *J. Chem. Phys.* **2004**, *121*, 681–690.

[199] Rodríguez, J. I.; Thompson, D. C.; Ayers, P. W.; Köster, A. M. Numerical Integration of Exchange-Correlation Energies and Potentials Using Transformed Sparse Grids. *J. Chem. Phys.* **2008**, *128*, 224103.

[200] Mulliken, R. S. Quelques Aspects De La Théorie Des Orbitales Moléculaires. *J. Chim. Phys.* **1949**, *46*, 497–542.

[201] Löwdin, P. Approximate Formulas for Many-Center Integrals in the Theory of Molecules and Crystals. *J. Chem. Phys.* **1953**, *21*, 374–375.

[202] Vahtras, O.; Almlöf, J.; Feyereisen, M. Integral Approximations for LCAO-SCF Calculations. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

[203] Ten-no, S.; Iwata, S. Three-Center Expansion of Electron Repulsion Integrals with Linear Combination of Atomic Electron Distributions. *Chem. Phys. Lett.* **1995**, *240*, 578–584.

[204] Weigend, F. A Fully Direct RI-HF Algorithm: Implementation, Optimised Auxiliary Basis Sets, Demonstration of Accuracy and Efficiency. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285–4291.

[205] Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of Approximate Integrals in Ab Initio Theory. an Application in MP2 Energy Calculations. *Chem. Phys. Lett.* **1993**, *208*, 359–363.

[206] Rendell, A. P.; Lee, T. J. Coupled-cluster Theory Employing Approximate Integrals: An Approach to Avoid the Input/output and Storage Bottlenecks. *J. Chem. Phys.* **1994**, *101*, 400–408.

[207] Dunlap, B. I. Robust and Variational Fitting. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2113–2116.

[208] Hättig, C.; Weigend, F. CC2 Excitation Energy Calculations on Large Molecules Using the Resolution of the Identity Approximation. *J. Chem. Phys.* **2000**, *113*, 5154.

[209] Grimme, S.; Waletzke, M. A Combination of Kohn–Sham Density Functional Theory and Multi-Reference Configuration Interaction Methods. *J. Chem. Phys.* **1999**, *111*, 5645–5655.

[210] Grimme, S.; Waletzke, M. Multi-Reference Møller–Plesset Theory: Computational Strategies for Large Molecules. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2075–2081.

[211] Werner, H.-J.; Manby, F. R.; Knowles, P. J. Fast Linear Scaling Second-Order Møller-Plesset Perturbation Theory (MP2) Using Local and Density Fitting Approximations. *J. Chem. Phys.* **2003**, *118*, 8149–8160.

[212] Schrader, D. M.; Prager, S. Use of Electrostatic Variation Principles in Molecular Energy Calculations. *J. Chem. Phys.* **1962**, *37*, 1456–1460.

[213] Baerends, E.; Ellis, D.; Ros, P. Self-Consistent Molecular Hartree—Fock—Slater Calculations I. the Computational Procedure. *Chem. Phys.* **1973**, *2*, 41–51.

[214] Whitten, J. L. Coulombic Potential Energy Integrals and Approximations. *J. Chem. Phys.* **1973**, *58*, 4496–4501.

[215] Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. on First-Row Diatomic Molecules and Local Density Models. *J. Chem. Phys.* **1979**, *71*, 4993.

[216] Billingsley, F. P.; Bloor, J. E. Limited Expansion of Diatomic Overlap (LEDO): A Near-Accurate Approximate Ab Initio LCAO MO Method. I. Theory and Preliminary Investigations. *J. Chem. Phys.* **1971**, *55*, 5178–5190.

[217] Gill, P. M. W.; Johnson, B. G.; Pople, J. A.; Taylor, S. W. Modeling the Potential of a Charge Distribution. *J. Chem. Phys.* **1992**, *96*, 7178–7179.

[218] Jung, Y.; Sodt, A.; Gill, P. M. W.; Head-Gordon, M. Auxiliary Basis Expansions for Large-Scale Electronic Structure Calculations. *Proc. Natl. Acad. Sci.* **2005**, *102*, 6692–6697.

[219] Sierka, M.; Hogekamp, A.; Ahlrichs, R. Fast Evaluation of the Coulomb Potential for Electron Densities Using Multipole Accelerated Resolution of Identity Approximation. *J. Chem. Phys.* **2003**, *118*, 9136–9148.

[220] Sodt, A.; Subotnik, J. E.; Head-Gordon, M. Linear Scaling Density Fitting. *J. Chem. Phys.* **2006**, *125*, 194109.

[221] Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules (International Series of Monographs on Chemistry)*; Oxford University Press, USA, 1994.

[222] Dirac, P. A. M. Note on Exchange Phenomena in the Thomas Atom. *Math. Proc. Cambridge Philos. Soc.* **1930**, *26*, 376–385.

[223] Treutler, O.; Ahlrichs, R. Efficient Molecular Numerical Integration Schemes. *J. Chem. Phys.* **1995**, *102*, 346–354.

[224] Jensen, F. *Introduction to Computational Chemistry*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006.

[225] te Velde, G.; Baerends, E. Numerical Integration for Polyatomic Systems. *J. Comput. Phys.* **1992**, *99*, 84–98.

[226] Gill, P. M.; Johnson, B. G.; Pople, J. A. A Standard Grid for Density Functional Calculations. *Chem. Phys. Lett.* **1993**, *209*, 506–512.

[227] Lindh, R.; Malmqvist, P.-&.; Gagliardi, L. Molecular Integrals by Numerical Quadrature. I. Radial Integration. *Theor. Chem. Accounts Theory, Comput. Model. (Theoretica Chim. Acta)* **2001**, *106*, 178–187.

[228] Lebedev, V. I. Quadratures on a Sphere. *USSR Comput. Maths Math. Phys.* **1976**, *16*, 10–24.

[229] Lebedev, V. I. Spherical Quadrature Formulas Exact to Orders 25-29. *Sib. Math. J.* **1977**, *18*, 99–107.

[230] Lebedev, V. I. Quadrature Formulas of Orders 41, 47, and 53 for the Sphere. *Russ. Acad. Sci. Dokl. Math.* **1992**, *45*, 587–592.

[231] Lebedev, V. I. A Quadrature Formula for the Sphere of 59th Algebraic Order of Accuracy. *Russ. Acad. Sci. Dokl. Math.* **1995**, *50*, 283–286.

[232] Lebedev, V. I.; Laikov, D. N. A Quadrature Formula for the Sphere of the 131st Algebraic Order of Accuracy. *Russ. Acad. Sci. Dokl. Math.* **1999**, *59*, 477–481.

[233] Murphy, K. P. *Mit Press. ISBN*; 2012.

[234] Álvarez, M. A.; Rosasco, L.; Lawrence, N. D. Kernels for Vector-Valued Functions: A Review. *Found. Trends. Mach. Learn.* **2012**, *4*, 195–266.

[235] Polson, N. G.; Sokolov, V. Deep Learning: A Bayesian Perspective. *Bayesian Anal.* **2017**, *12*, 1275–1304.

[236] Rasmussen, C. E.; Williams, C. K. I. *Gaussian Process. Mach. Learn.*; The MIT Press, 2005.

[237] MacKay, D. J. C. Bayesian Interpolation. *Neural Comput.* **1992**, *4*, 415–447.

[238] Kruschke, J. K. Bayesian Data Analysis. *Wiley Interdiscip. Rev. Cogn. Sci.* **2010**, *1*, 658–676.

[239] MacKay, D. Information Theory, Inference, and Learning Algorithms. *IEEE Trans. Inf. Theory* **2004**, *50*, 2544–2545.

[240] Dudley, R. M. *Real Anal. Probab.*; Chapman and Hall/CRC, 2018.

[241] Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel Methods in Machine Learning. *Ann. Stat.* **2008**, *36*, 1171–1220.

[242] Schölkopf, B.; Smola, A. J. *Proc. 2002 Int. Conf. Mach. Learn. Cybern.*; The MIT Press, 2018.

[243] Park, K. I. *Fundam. Probab. Stoch. Process. with Appl. to Commun.*; Springer International Publishing: Cham, 2018.

[244] Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

[245] Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, 095003.

[246] Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B* **2014**, *89*, 205118.

[247] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; 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.

[248] Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.

[249] Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094–1101.

[250] Huang, B.; von Lilienfeld, O. A. Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *J. Chem. Phys.* **2016**, *145*, 161102.

[251] Huang, B.; von Lilienfeld, O. A. The "DNA" of Chemistry: Scalable Quantum Machine Learning with "amons". *Preprint* **2017**, *arXiv:1707.04146*.

[252] Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, 1st ed.; Dover Publications, Inc.: Mineola, 1996.

[253] Kohn, W. Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms. *Phys. Rev. Lett.* **1996**, *76*, 3168–3171.

[254] Prodan, E.; Kohn, W. Nearsightedness of Electronic Matter. *Proc. Nati. Acad. Sci. USA* **2005**, *102*, 11635–11638.

[255] Huang, B.; Symonds, N. O.; von Lilienfeld, O. A. The Fundamentals of Quantum Machine Learning. **2018**, *arXiv:1807.04259*.

[256] Glielmo, A.; Sollich, P.; De Vita, A. Accurate Interatomic Force Fields Via Machine Learning with Covariant Kernels. *Phys. Rev. B* **2017**, *95*, 214302.

[257] Grisafi, A.; Wilkins, D. M.; Willatt, M. J.; Ceriotti, M. *Atomic-Scale Representation and Statistical Learning of Tensorial Properties*; 2019; pp 1–21.

[258] Liang, C.; Tocci, G.; Wilkins, D. M.; Grisafi, A.; Roke, S.; Ceriotti, M. Solvent Fluctuations and Nuclear Quantum Effects Modulate the Molecular Hyperpolarizability of Water. *Phys. Rev. B* **2017**, *96*, 041407.

[259] Weinert, U. Spherical Tensor Representation. *Arch. Ration. Mech. Anal.* **1980**, *74*, 165–196.

[260] Wigner, E. P.; Fano, U. Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra. *Am. J. Phys.* **1960**, *28*, 408–409.

[261] Biedenharn, L. C.; Louck, J. D.; Carruthers, P. A. *Angular Momentum Quantum Phys.*; Cambridge University Press, 1984.

[262] Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A Look at the Density Functional Theory Zoo with the Advanced GMTKN55 Database for General Main Group Thermochemistry, Kinetics and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

[263] Řezáč, J.; Hobza, P. Benchmark Calculations of Interaction Energies in Noncovalent Complexes and Their Applications. *Chem. Rev.* **2016**, *116*, 5038–5071.

[264] Müller-Dethlefs, K.; Hobza, P. Noncovalent Interactions: a Challenge for Experiment and Theory. *Chem. Rev.* **2000**, *100*, 143–168.

[265] Stöhr, M.; Van Voorhis, T.; Tkatchenko, A. Theory and Practice of Modeling Van Der Waals Interactions in Electronic-Structure Calculations. *Chemical Society Reviews* **2019**, *48*, 4118–4154.

[266] Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNa Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

[267] Podeszwa, R.; Patkowski, K.; Szalewicz, K. Improved Interaction Energy Benchmarks for Dimers of Biological Relevance. *Phys. Chem. Chem. Phys.* **2010**, *12*, 5974–5979.

[268] Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. Basis Set Consistent Revision of the S22 Test Set of Noncovalent Interaction Energies. *J. Chem. Phys.* **2010**, *132*, 144104.

[269] Molnar, L. F.; He, X.; Wang, B.; Merz, K. M. Further Analysis and Comparative Study of Intermolecular Interactions Using Dimers from the S22 Database. *J. Chem. Phys.* **2009**, *131*, 08B603.

[270] Gráfová, L.; Pitoňák, M.; Řezáč, J.; Hobza, P. Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set. *J. Chem. Theory Comput.* **2010**, *6*, 2365–2376.

[271] Řezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.

[272] Risthaus, T.; Grimme, S. Benchmarking of London Dispersion-Accounting Density Functional Theory Methods on Very Large Molecular Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 1580–91.

[273] Burns, L. A.; Vázquez-Mayagoitia, Á.; Sumpter, B. G.; Sherrill, C. D. Density-Functional Approaches to Noncovalent Interactions: A Comparison of Dispersion Corrections (DFT-D), Exchange-Hole Dipole Moment (XDM) Theory, and Specialized Functionals. *J. Chem Phys.* **2011**, *134*, 084107.

[274] Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis Set Convergence of the Coupled-Cluster Correction, $\delta_{MP2}^{CCSD(T)}$: Best Practices for Benchmarking Non-Covalent Interactions and the Attendant Revision of the S22, NBC10, HBC6, and HSG Databases. *J. Chem. Phys.* **2011**, *135*, 194102.

[275] Smith, D. G. A.; Burns, L. A.; Patkowski, K.; Sherrill, C. D. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *J. Phys. Chem. Lett.* **2016**, *7*, 2197–2203.

[276] Goerigk, L.; Grimme, S. A General Database for Main Group Thermochemistry, Kinetics, and Noncovalent Interactions - Assessment of Common and Reparameterized (meta-) GGA Density Functionals. *J. Chem. Theory Comput.* **2010**, *6*, 107–126.

[277] Goerigk, L.; Grimme, S. A Thorough Benchmark of Density Functional Methods for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.

[278] Schneebeli, S. T.; Bochevarov, A. D.; Friesner, R. A. Parameterization of a B3LYP Specific Correction for Noncovalent Interactions and Basis Set Superposition Error on a Gigantic

Data Set of CCSD(T) Quality Noncovalent Interaction Energies. *J. Chem. Theory Comput.* **2011**, *7*, 658–668.

[279] McGibbon, R. T.; Taube, A. G.; Donchev, A. G.; Siva, K.; Hernández, F.; Hargus, C.; Law, K.-H.; Klepeis, J. L.; Shaw, D. E. Improving the Accuracy of Møller-Plesset Perturbation Theory with Neural Networks. *J. Chem. Phys.* **2017**, *147*, 161725.

[280] Mardirossian, N.; Head-Gordon, M. *ω* B97M-V: A Combinatorially Optimized, Range-Separated Hybrid, Meta-GGA Density Functional with VV10 Nonlocal Correlation. *J. Chem. Phys.* **2016**, *144*, 214110.

[281] Hains, A. W.; Liang, Z.; Woodhouse, M. A.; Gregg, B. A. Molecular Semiconductors in Organic Photovoltaic Cells. *Chem. Rev.* **2010**, *110*, 6689–6735.

[282] Lüssem, B.; Keum, C.-M.; Kasemann, D.; Naab, B.; Bao, Z.; Leo, K. Doped Organic Transistors. *Chem. Rev.* **2016**, *116*, 13714–13751.

[283] Ostroverkhova, O. Organic Optoelectronic Materials: Mechanisms and Applications. *Chem. Rev.* **2016**, *116*, 13279–13412.

[284] Walzer, K.; Maennig, B.; Pfeiffer, M.; Leo, K. Highly Efficient Organic Devices Based on Electrically Doped Transport Layers. *Chem. Rev.* **2007**, *107*, 1233–71.

[285] Coropceanu, V.; Cornil, J.; da Silva Filho, D. A.; Olivier, Y.; Silbey, R.; Brédas, J.-L. Charge Transport in Organic Semiconductors. *Chem. Rev.* **2007**, *107*, 926–952.

[286] Corminboeuf, C. Minimizing Density Functional Failures for Non-Covalent Interactions Beyond van der Waals Complexes. *Acc. Chem. Res.* **2014**, *47*, 3217–3224.

[287] Otero-de-la Roza, A.; Johnson, E. R.; DiLabio, G. A. Halogen Bonding from Dispersion-Corrected Density-Functional Theory: The Role of Delocalization Error. *J. Chem. Theory Comput.* **2014**, *10*, 5436–5447.

[288] Boese, A. D. Density Functional Theory and Hydrogen Bonds: Are We There Yet? *ChemPhysChem* **2015**, *16*, 978–985.

[289] Bauzá, A.; Alkorta, I.; Frontera, A.; Elguero, J. on the Reliability of Pure and Hybrid DFT Methods for the Evaluation of Halogen, Chalcogen, and Pnicogen Bonds Involving Anionic and Neutral Electron Donors. *J. Chem. Theory Comput.* **2013**, *9*, 5201–5210.

[290] Soniat, M.; Rogers, D. M.; Rempe, S. B. Dispersion- and Exchange-Corrected Density Functional Theory for Sodium Ion Hydration. *J. Chem. Theory Comput.* **2015**, *11*, 2958–2967.

[291] Shi, R.; Huang, X.; Su, Y.; Lu, H.-G.; Li, S.-D.; Tang, L.; Zhao, J. Which Density Functional Should Be Used to Describe Protonated Water Clusters? *J. Phys. Chem. A* **2017**, *121*, 3117–3127.

[292] Schreiner, P. R. Relative Energy Computations with Approximate Density Functional Theory—a Caveat! *Angew. Chemie Int. Ed.* **2007**, *46*, 4217–4219.

[293] Grimme, S.; Steinmetz, M.; Korth, M. Stereoelectronic Substituent Effects in Saturated Main Group Molecules: Severe Problems of Current Kohn-Sham Density Functional Theory. *J. Chem. Theory Comput.* **2007**, *3*, 42–45.

[294] Facchetti, A. π-Conjugated Polymers for Organic Electronics and Photovoltaic Cell Applications. *Chem. Mater.* **2011**, *23*, 733–758.

[295] Wang, C.; Dong, H.; Hu, W.; Liu, Y.; Zhu, D. Semiconducting π-Conjugated Systems in Field-Effect Transistors: A Material Odyssey of Organic Electronics. *Chem. Rev.* **2012**, *112*, 2208–2267.

[296] Wang, C.; Dong, H.; Jiang, L.; Hu, W. Organic Semiconductor Crystals. *Chem. Soc. Rev.* **2018**, *47*, 422–500.

[297] Gryn'ova, G.; Nicolaï, A.; Prlj, A.; Ollitrault, P.; Andrienko, D.; Corminboeuf, C. Charge Transport in Highly Ordered Organic Nanofibrils: Lessons from Modelling. *J. Mater. Chem. C* **2017**, *5*, 350–361.

[298] Chai, J.-D. D.; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128*, 084106.

[299] Lin, Y.-S.; Tsai, C.-W.; Li, G.-D.; Chai, J.-D. Long-Range Corrected Hybrid Meta-Generalized-Gradient Approximations with Dispersion Corrections. *J. Chem. Phys.* **2012**, *136*, 154109.

[300] Mardirossian, N.; Head-Gordon, M. ωB97X-V: A 10-Parameter, Range-Separated Hybrid, Generalized Gradient Approximation Density Functional with Nonlocal Correlation, Designed by a Survival-Of-The-Fittest Strategy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904.

[301] Anthony, J. E.; Eaton, D. L.; Parkin, S. R. A Road Map to Stable, Soluble, Easily Crystallized Pentacene Derivatives. *Org. Lett.* **2002**, *4*, 15–18.

[302] Herwig, P.; Kayser, C. W.; Müllen, K.; Spiess, H. W. Columnar Mesophases of Alkylated Hexa-Peri-Hexabenzocoronenes with Remarkably Large Phase Widths. *Adv. Mater.* **1996**, *8*, 510–513.

[303] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

[304] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Chem. Phys.* **1994**, *98*, 11623–11627.

[305] Frisch, M. J. et al. Gaussian16 Revision B.01. 2016.

[306] Mas-Torrent, M.; Hadley, P.; Bromley, S. T.; Ribas, X.; Tarrés, J.; Mas, M.; Molins, E.; Veciana, J.; Rovira, C. Correlation Between Crystal Structure and Mobility in Organic Field-Effect Transistors Based on Single Crystals of Tetrathiafulvalene Derivatives. *J. Am. Chem. Soc* **2004**, *126*, 8546–8553.

[307] Li, X.-C.; Sirringhaus, H.; Garnier, F.; Holmes, A. B.; Moratti, S. C.; Feeder, N.; Clegg, W.; Teat, S. J.; Friend, R. H. A Highly π-Stacked Organic Semiconductor for Thin-Film Transistors Based on Fused Thiophenes. *J. Am. Chem. Soc.* **1998**, *120*, 2206–2207.

[308] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. B* **2016**, *72*, 171–179.

[309] Afonina, I.; Skabara, P. J.; Vilela, F.; Kanibolotsky, A. L.; Forgie, J. C.; Bansal, A. K.; Turnbull, G. A.; Samuel, I. D.; Labram, J. G.; Anthopoulos, T. D.; Coles, S. J.; Hursthouse, M. B. Synthesis and Characterisation of New Diindenodithienothiophene (DITT) Based Materials. *J. Mater. Chem.* **2010**, *20*, 1112–1116.

[310] Chen, M.-C.; Vegiraju, S.; Huang, C.-M.; Huang, P.-Y.; Prabakaran, K.; Yau, S. L.; Chen, W.-C.; Peng, W.-T.; Chao, I.; Kim, C.; Tao, Y.-T. Asymmetric Fused Thiophenes for Field-Effect Transistors: Crystal Structure–film Microstructure–transistor Performance Correlations. *J. Mater. Chem. C* **2014**, *2*, 8892–8902.

[311] Ebata, H.; Miyazaki, E.; Yamamoto, T.; Takimiya, K. Synthesis, Properties, and Structures of Benzo[1,2-B:4,5-B']bis[b] Benzothiophene and Benzo[1,2-B:4,5-B']bis[b]benzoselenophene. *Org. Lett.* **2007**, *9*, 4499–4502.

[312] Wang, C.; Dong, H.; Li, H.; Zhao, H.; Meng, Q.; Hu, W. Dibenzothiophene Derivatives: From Herringbone to Lamellar Packing Motif. *Cryst. Growth Des.* **2010**, *10*, 4155–4160.

[313] Antolini, L.; Horowitz, G.; Kouki, F.; Garnier, F. Polymorphism in Oligothiophenes with an Even Number of Thiophene Subunits. *Adv.Mater.* **1998**, *10*, 382–385.

[314] Yamazaki, D.; Nishinaga, T.; Komatsu, K. Radical Cation of Dibenzothiophene Fully Annelated with Bicyclo[2.2.2]octene Units: X-Ray Crystal Structure and Electronic Properties. *Org. Lett.* **2004**, *6*, 4179–4182.

[315] Zhang, X.; Johnson, J. P.; Kampf, J. W.; Matzger, A. J. Ring Fusion Effects on the Solid-State Properties of α- Oligothiophenes. *Chem. Mater.* **2006**, *18*, 3470–3476.

[316] Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural Triple Excitations in Local Coupled Cluster Calculations with Pair Natural Orbitals. *J Chem. Phys.* **2013**, *139*, 134101.

[317] Riplinger, C.; Neese, F. An Efficient and Near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138*, 034106.

[318] Neese, F. The ORCA Program System. *Wiley Interdiscip. Rev. Comput. Mol. Sci* **2012**, *2*, 73–78.

[319] Zhang, J.; Dolg, M. Dispersion Interaction Stabilizes Sterically Hindered Double Fullerenes. *Chem. Eur. J.* **2014**, *20*, 13909–13912.

[320] Minenkov, Y.; Chermak, E.; Cavallo, L. Accuracy of DLPNO-CCSD(T) Method for Noncovalent Bond Dissociation Enthalpies from Coinage Metal Cation Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 4664–4676.

[321] Neese, F.; Valeev, E. F. Revisiting the Atomic Natural Orbital Approach for Basis Sets: Robust Systematic Basis Sets for Explicitly Correlated and Conventional Correlated Ab Initio Methods? *J. Chem. Theory Comput.* **2011**, *7*, 33–43.

[322] Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. the Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

[323] Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19*, 553–566.

[324] Saitow, M.; Becker, U.; Riplinger, C.; Valeev, E. F.; Neese, F. A New Near-Linear Scaling, Efficient and Accurate, Open-Shell Domain-Based Local Pair Natural Orbital Coupled Cluster Singles and Doubles Theory. *J. Chem. Phys.* **2017**, *146*, 164105.

[325] 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. Che. Chem. Phys.* **2005**, *7*, 3297–305.

[326] Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Non-Empirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.

[327] Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

[328] Gordon, M. S.; Schmidt, M. W. In *Theory and Applications of Computational Chemistry*; Dykstra, C., Frenking, G., Kim, K., Scuseria, G., Eds.; Elsevier: Amsterdam, 2005.

[329] Shao, Y. et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.

[330] Ren, X.; Rinke, P.; Joas, C.; Scheffler, M. Random-Phase Approximation and Its Applications in Computational Chemistry and Materials Science. *J. Mat. Sci.* **2012**, *47*, 7447–7471.

[331] Eshuis, H.; Bates, J. E.; Furche, F. Electron Correlation Methods Based on the Random Phase Approximation. *Theor. Chem. Acc.* **2012**, *131*, 1084.

[332] Eshuis, H.; Yarkony, J.; Furche, F. Fast Computation of Molecular Random Phase Approximation Correlation Energies Using Resolution of the Identity and Imaginary Frequency Integration. *J. Chem. Phys.* **2010**, *132*, 234114.

[333] Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. Turbomole. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4*, 91–100.

[334] Perdew, J. J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

[335] Perdew, J. P.; Burke, K.; Ernzerhof, M. Erratum: Generalized Gradient Approximation Made Simple (Physical Review Letters (1996) 77 (3865)). *Phys. Rev. Lett.* **1997**, *78*, 1396.

[336] Del Ben, M.; Hutter, J.; Vandevondele, J. Electron Correlation in the Condensed Phase from a Resolution of Identity Approach Based on the Gaussian and Plane Waves Scheme. *J. Chem. Theory Comput.* **2013**, *9*, 2654–2671.

[337] Řezáč, J.; Hobza, P. Describing Noncovalent Interactions Beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) at the Complete Basis Set Limit? *J. Chem. Theory Comput.* **2013**, *9*, 2151–2155.

[338] Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. Molpro: A General-purpose Quantum Chemistry Program Package. *WIREs Comp. Mol. Sci.* **2012**, *2*, 242–253.

[339] Gonthier, J. F.; Sherrill, C. D. Density-Fitted Open-Shell Symmetry-Adapted Perturbation Theory and Application to π-Stacking in Benzene Dimer Cation and Ionized DNa Base Pair Steps. *J. Chem. Phys.* **2016**, *145*, 134106.

[340] Parrish, R. M. et al. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**, *13*, 3185–3197.

[341] Parker, T. M.; Burns, L. A.; Parrish, R. M.; Ryno, A. G.; Sherrill, C. D. Levels of Symmetry Adapted Perturbation Theory (SAPT). I. Efficiency and Performance for Interaction Energies. *J. Chem. Phys.* **2014**, *140*, 094106.

[342] Becke, A. D. Density-Functional Thermochemistry. V. Systematic Optimization of Exchange-Correlation Functionals. *J. Chem. Phys.* **1997**, *107*, 8554.

[343] Dobson, J. F. Beyond Pairwise Additivity in London Dispersion Interactions. *Int. J. Quant. Chem.* **2014**, *114*, 1157–1161.

[344] Axilrod, B. M.; Teller, E. Interaction of the Van Der Waals Type Between Three Atoms. *J. Chem. Phys.* **1943**, *11*, 299–300.

[345] Gobre, V. V.; Tkatchenko, A. Scaling Laws for Van Der Waals Interactions in Nanostructured Materials. *Nat. Comm.* **2013**, *4*, 2341.

[346] Ruzsinszky, A.; Perdew, J. P.; Tao, J.; Csonka, G. I.; Pitarke, J. M. Van Der Waals Coefficients for Nanostructures: Fullerenes Defy Conventional Wisdom. *Phys. Rev. Lett.* **2012**, *109*, 233203.

[347] Ambrosetti, A.; Alfè, D.; DiStasio, R. A.; Tkatchenko, A. Hard Numbers for Large Molecules: Toward Exact Energetics for Supramolecular Systems. *J. Phys. Chem. Lett.* **2014**, *5*, 849–855.

[348] Otero-de-la Roza, A.; Johnson, E. R. Predicting Energetics of Supramolecular Systems Using the XDM Dispersion Model. *J. Chem. Theory Comput.* **2015**, *11*, 4033–40.

[349] Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208.

[350] Furche, F.; Van Voorhis, T. Fluctuation-Dissipation Theorem Density-Functional Theory. *J. Chem. Phys.* **2005**, *122*, 164106.

[351] Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. Long-Range-Corrected Hybrids Including Random Phase Approximation Correlation. *J. Chem. Phys.* **2009**, *130*, 081105.

[352] Toulouse, J.; Gerber, I. C.; Jansen, G.; Savin, A.; Ángyán, J. G. Adiabatic-Connection Fluctuation-Dissipation Density-Functional Theory Based on Range Separation. *Phys. Rev. Lett.* **2009**, *102*, 096404.

[353] Zhu, W.; Toulouse, J.; Savin, A.; Ángyán, J. G. Range-Separated Density-Functional Theory with Random Phase Approximation Applied to Noncovalent Intermolecular Interactions. *J. Chem. Phys.* **2010**, *132*, 244108.

[354] Toulouse, J.; Zhu, W.; Savin, A.; Jansen, G.; Ángyán, J. G. Closed-Shell Ring Coupled Cluster Doubles Theory with Range Separation Applied on Weak Intermolecular Interactions. *J. Chem. Phys.* **2011**, *135*, 084119.

[355] Harl, J.; Kresse, G. Cohesive Energy Curves for Noble Gas Solids Calculated by Adiabatic Connection Fluctuation-Dissipation Theory. *Phys. Rev. B* **2008**, *77*, 045136.

[356] Lu, D.; Li, Y.; Rocca, D.; Galli, G. Ab Initio Calculation of Van Der Waals Bonded Molecular Crystals. *Phys. Rev. Lett.* **2009**, *102*, 206411.

[357] Li, Y.; Lu, D.; Nguyen, H. V.; Galii, G. Van Der Waals Interactions in Molecular Assemblies from First-Principles Calculations. *J. Phys. Chem. A* **2010**, *114*, 1944–1952.

[358] Scuseria, G. E.; Henderson, T. M.; Sorensen, D. C. The Ground State Correlation Energy of the Random Phase Approximation from a Ring Coupled Cluster Doubles Approach. *J. Chem. Phys.* **2008**, *129*, 231101.

[359] Sinnokrot, M. O.; Sherrill, C. D. Substituent Effects in π-π Interactions: Sandwich and T-Shaped Configurations. *J. Am. Chem. Soc.* **2004**, *126*, 7690–7697.

[360] Sherrill, C. D. Energy Component Analysis of π Interactions. *Acc. Chem. Res.* **2013**, *46*, 1020–1028.

[361] Steinmann, S. N.; Csonka, G.; Corminboeuf, C. Unified Inter- and Intramolecular Dispersion Correction Formula for Generalized Gradient Approximation Density Functional Theory. *J. Chem. Theory Comput.* **2009**, *5*, 2950–2958.

[362] Schmidt, J.; Benavides-Riveros, C. L.; Marques, M. A. L. Machine Learning the Physical Nonlocal Exchange–Correlation Functional of Density-Functional Theory. *J. Phys. Chem. Lett.* **2019**, *10*, 6425–6431.

[363] Wagner, J. P.; Schreiner, P. R. London Dispersion in Molecular Chemistry-Reconsidering Steric Effects. *Angew. Chem. Int. Ed.* **2015**, *54*, 12274–12296.

[364] Chalasinski, G.; Szczesniak, M. M. State of the Art and Challenges of the Ab Initio Theory of Intermolecular Interactions. *Chem. Rev.* **2000**, *100*, 4227–4252.

[365] Klimeš, J.; Michaelides, A. Perspective: Advances and Challenges in Treating Van Der Waals Dispersion Forces in Density Functional Theory. *J. Chem. Phys.* **2012**, *137*, 120901.

[366] Zhao, Y.; Truhlar, D. G. A New Local Density Functional for Main-Group Thermochemistry, Transition Metal Bonding, Thermochemical Kinetics, and Noncovalent Interactions. *J. Chem. Phys.* **2006**, *125*, 194101.

[367] Cheng, P.; Zhan, X. Stability of Organic Solar Cells: Challenges and Strategies. *Chem. Soc. Rev.* **2016**, *45*, 2544–2582.

[368] Trouillas, P.; Sancho-García, J. C.; De Freitas, V.; Gierschner, J.; Otyepka, M.; Dangles, O. Stabilizing and Modulating Color by Copigmentation: Insights from Theory and Experiment. *Chem. Rev.* **2016**, *116*, 4937–4982.

[369] Luisier, N.; Ruggi, A.; Steinmann, S. N.; Favre, L.; Gaeng, N.; Corminboeuf, C.; Severin, K. A Ratiometric Fluorescence Sensor for Caffeine. *Org. Biomol. Chem.* **2012**, *10*, 7487.

[370] Shinohara, H.; Nishi, N. Excited State Lifetimes and Appearance Potentials of Benzene Dimer and Trimer. *J. Chem. Phys.* **1989**, *91*, 6743.

[371] Saigusa, H.; Limt, E. C. Excited-State Dynamics of Aromatic Clusters: Correlation Between Exciton Interactions and Excimer Formation Dynamics. *J. Phys. Chem.* **1995**, *99*, 15738–15747.

[372] Diri, K.; Krylov, A. I. Electronic States of the Benzene Dimer: A Simple Case of Complexity. *J. Phys. Chem. A* **2012**, *116*, 653–662.

[373] Dreuw, A.; Head-Gordon, M. Single-Reference Ab Initio Methods for the Calculation of Excited States of Large Molecules. *Chem. Rev.* **2005**, *105*, 4009–4037.

[374] Komasa, J. in Search for the Negative Polarizability States – the EF 1Σg+ State of Hydrogen Molecule. *Adv. Quantum Chem.* **2005**, *48*, 151–159.

[375] Milton, K. A.; Parashar, P.; Pourtolami, N.; Brevik, I. Casimir-Polder Repulsion: Polarizable Atoms, Cylinders, Spheres, and Ellipsoids. *Phys. Rev. D - Part. Fields, Gravit. Cosmol.* **2012**, *85*, 025008.

[376] Milton, K. A.; Abalo, E. K.; Parashar, P.; Pourtolami, N.; Brevik, I.; Ellingsen, S. A. Repulsive Casimir and Casimir-Polder Forces. *J. Phys. A Math. Theor.* **2012**, *45*, 374006.

[377] Power, E. A.; Thirunamachandran, T. Dispersion Forces Between Molecules with One or Both Molecules Excited. *Phys. Rev. A* **1995**, *51*, 3660–3666.

[378] Power, E. A.; Thirunamachandran, T. Two- and Three-Body Dispersion Forces with One Excited Molecule. *Chem. Phys.* **1995**, *198*, 5–17.

[379] Sherkunov, Y. Van Der Waals Interaction of Excited Media. *Phys. Rev. A* **2005**, *72*, 052703.

[380] Barcellona, P.; Passante, R.; Rizzuto, L.; Buhmann, S. Y. Van Der Waals Interactions Between Excited Atoms in Generic Environments. *Phys. Rev. A* **2016**, *94*, 012705.

[381] Adhikari, C. M.; Debierre, V.; Matveev, A.; Kolachevsky, N.; Jentschura, U. D. Long-Range Interactions of Hydrogen Atoms in Excited States. I. 2S-1s Interactions and Dirac- δ Perturbations. *Phys. Rev. A* **2017**, *95*, 022703.

[382] Jentschura, U. D.; Debierre, V. Long-Range Tails in Van Der Waals Interactions of Excited-State and Ground-State Atoms. *Phys. Rev. A* **2017**, *95*, 042506.

[383] Adelman, S. A.; Szabo, A. Coulomb Approximation for Multipole Polarizabilities and Dispersion Forces: Analytic Static Polarizabilities of Ground and Excited State Atoms. *J. Chem. Phys.* **1973**, *58*, 687–696.

[384] Norman, P.; Jiemchooroj, A.; Sernelius, B. E. First Principle Calculations of Dipole-Dipole Dispersion Coefficients for the Ground and First $\pi \rightarrow \pi^*$ Excited States of Some Azabenzenes. *J. Comput. Methods Sci. Eng.* **2004**, *4*, 321–332.

[385] Yan, Z.-C.; Babb, J.; Dalgarno, A.; Drake, G. W. F. Variational Calculations of Dispersion Coefficients for Interactions Among H, He, and Li Atoms. *Phys. Rev. A* **1996**, *54*, 2824–2833.

[386] Gritsenko, O.; Baerends, E. J. A Simple Natural Orbital Mechanism of "pure" Van Der Waals Interaction in the Lowest Excited Triplet State of the Hydrogen Molecule. *J. Chem. Phys.* **2006**, *124*, 054115.

[387] Amicangelo, J. C. Theoretical Study of the Benzene Excimer Using Time-Dependent Density Functional Theory. *J. Phys. Chem. A* **2005**, *109*, 9174–9182.

[388] Huenerbein, R.; Grimme, S. Time-Dependent Density Functional Study of Excimers and Exciplexes of Organic Molecules. *Chem. Phys.* **2008**, *343*, 362–371.

[389] Barone, V.; Biczysko, M.; Pavone, M. The Role of Dispersion Correction to DFT for Modelling Weakly Bound Molecular Complexes in the Ground and Excited Electronic States. *Chem. Phys.* **2008**, *346*, 247–256.

[390] Briggs, E. A.; Besley, N. A. Modelling Excited States of Weakly Bound Complexes with Density Functional Theory. *Phys. Chem. Chem. Phys.* **2014**, *16*, 14455–14462.

[391] Ikabata, Y.; Nakai, H. Extension of Local Response Dispersion Method to Excited-State Calculation Based on Time-Dependent Density Functional Theory. *J. Chem. Phys.* **2012**, *137*, 124106.

[392] Gilbert, A. T. B.; Besley, N. A.; Gill, P. M. W. Self-Consistent Field Calculations of Excited States Using the Maximum Overlap Method (MOM). *J. Phys. Chem. A* **2008**, *112*, 13164–13171.

[393] Ershova, O. V.; Besley, N. A. Can Density Functional Theory Describe the NO($X^2\Pi$)-Ar and NO($A^2\Sigma^+$)-Ar Van Der Waals Complexes? *J. Chem. Phys.* **2012**, *136*, 244313.

[394] Schweighauser, L.; Strauss, M. A.; Bellotto, S.; Wegner, H. A. Attraction or Repulsion? London Dispersion Forces Control Azobenzene Switches. *Angew. Chem. Int. Ed.* **2015**, *54*, 13436–13439.

[395] Irie, M. Diarylethenes for Memories and Switches. *Chem. Rev.* **2000**, *100*, 1685–1716.

[396] Irie, M.; Fukaminato, T.; Matsuda, K.; Kobatake, S. Photochromism of Diarylethene Molecules and Crystals: Memories, Switches, and Actuators. *Chem. Rev.* **2014**, *114*, 12174–12277.

[397] Waldeck, D. H. Photoisomerization Dynamics of Stilbenes. *Chem. Rev.* **1991**, *91*, 415–436.

[398] Minezawa, N.; Gordon, M. S. Photoisomerization of Stilbene: A Spin-Flip Density Functional Theory Approach. *J. Phys. Chem. A* **2011**, *115*, 7901–7911.

[399] Ioffe, I. N.; Granovsky, A. A. Photoisomerization of Stilbene: The Detailed XMCQDPT2 Treatment. *J. Chem. Theory Comput.* **2013**, *9*, 4973–4990.

[400] Mallory, F. B.; Mallory, C. W. *Organic Reactions*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1984; Vol. 30; pp 1–456.

[401] Hirata, S.; Head-Gordon, M. Time-Dependent Density Functional Theory Within the Tamm–Dancoff Approximation. *Chem. Phys. Lett.* **1999**, *314*, 291–299.

[402] Tapavicza, E.; Tavernelli, I.; Rothlisberger, U.; Filippi, C.; Casida, M. E. Mixed Time-Dependent Density-Functional Theory/classical Trajectory Surface Hopping Study of Oxirane Photochemistry. *J. Chem. Phys.* **2008**, *129*, 124108.

[403] Adamo, C.; Barone, V. Toward Reliable Density Functional Methods Without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110*, 6158.

[404] Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew–Burke–Ernzerhof Exchange-Correlation Functional. *J. Chem. Phys.* **1999**, *110*, 5029.

[405] Yanai, T.; Tew, D. P.; Handy, N. C. A New Hybrid Exchange-Correlation Functional Using the Coulomb-Attenuating Method (CAM-B3LYP). *Chemical Physics Letters* **2004**, *393*, 51–57.

[406] Christiansen, O.; Koch, H.; Jorgensen, P. The Second-Order Approximate Coupled Cluster Singles and Doubles Model CC2. *Chem. Phys. Lett.* **1995**, *243*, 409–418.

[407] Tuna, D.; Lefrancois, D.; Wolański, Ł.; Gozem, S.; Schapiro, I.; Andruniów, T.; Dreuw, A.; Olivucci, M. Assessment of Approximate Coupled-Cluster and Algebraic-Diagrammatic-Construction Methods for Ground- and Excited-State Reaction Paths and the Conical-Intersection Seam of a Retinal-Chromophore Model. *J. Chem. Theory Comput.* **2015**, *11*, 5758–5781.

[408] Ufimtsev, I. S.; Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5*, 2619–2628.

[409] Isborn, C. M.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J. Excited-State Electronic Structure with Configuration Interaction Singles and Tamm-Dancoff Time-Dependent Density Functional Theory on Graphical Processing Units. *J. Chem. Theory Comput.* **2011**, *7*, 1814–1823.

[410] Titov, A. V.; Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. Generating Efficient Quantum Chemistry Codes for Novel Architectures. *J. Chem. Theory Comput.* **2013**, *9*, 213–221.

[411] Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

[412] Koritsanszky, T. S.; Coppens, P. Chemical Applications of X-Ray Charge-Density Analysis. *Chem. Rev.* **2001**, *101*, 1583–1628.

[413] Gatti, C., Macchi, P., Eds. *Modern Charge-Density Analysis*, 1st ed.; Springer Netherlands, 2012.

[414] Meyer, J. C.; Kurasch, S.; Park, H. J.; Skakalova, V.; Künzel, D.; Groß, A.; Chuvilin, A.; Algara-Siller, G.; Roth, S.; Iwasaki, T.; Starke, U.; Smet, J. H.; Kaiser, U. Experimental Analysis of Charge Redistribution Due to Chemical Bonding by High-Resolution Transmission Electron Microscopy. *Nat. Mater.* **2011**, *10*, 209–215.

[415] Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8*, 6924–6935.

[416] Fowler, A. T.; Pickard, C. J.; Elliott, J. A. Managing Uncertainty in Data-Derived Densities to Accelerate Density Functional Theory. *J. Phys. Mater.* **2019**, *2*, 034001.

[417] Foster, J. M.; Boys, S. F. Canonical Configurational Interaction Procedure. *Rev. Mod. Phys.* **1960**, *32*, 300–302.

[418] Edmiston, C.; Ruedenberg, K. Localized Atomic and Molecular Orbitals. *Rev. Mod. Phys.* **1963**, *35*, 457–464.

[419] Pipek, J.; Mezey, P. G. A Fast Intrinsic Localization Procedure Applicable for Ab-Initio and Semiempirical Linear Combination of Atomic Orbital Wave Functions. *J. Chem. Phys.* **1989**, *90*, 4916–4926.

[420] Marzari, N.; Vanderbilt, D. Maximally Localized Generalized Wannier Functions for Composite Energy Bands. *Phys. Rev. B* **1997**, *56*, 12847–12865.

[421] Høyvik, I.-M.; Jansik, B.; Jørgensen, P. Orbital Localization Using Fourth Central Moment Minimization. *J. Chem. Phys.* **2012**, *137*, 224114.

[422] Dominiak, P. M.; Volkov, A.; Li, X.; Messerschmidt, M.; Coppens, P. A Theoretical Databank of Transferable Aspherical Atoms and Its Application to Electrostatic Interaction Energy Calculations of Macromolecules. *J. Chem. Theory Comput.* **2007**, *3*, 232–247.

[423] Pichon-Pesme, V.; Lecomte, C.; Lachekar, H. On Building a Data Bank of Transferable Experimental Electron Density Parameters Applicable to Polypeptides. *J. Phys. Chem.* **1995**, *99*, 6242–6250.

[424] Guillot, B.; Jelsch, C.; Podjarny, A.; Lecomte, C. Charge-Density Analysis of a Protein Structure at Subatomic Resolution: The Human Aldose Reductase Case. *Acta Crystallogr., Sect. D* **2008**, *64*, 567–588.

[425] Ghermani, N.-E.; Bouhmaida, N.; Lecomte, C. Modelling Electrostatic Potential from Experimentally Determined Charge Densities. I. Spherical-Atom Approximation. *Acta Crystallogr., Sect. A* **1993**, *49*, 781–789.

[426] Bouhmaida, N.; Ghermani, N.-E.; Lecomte, C.; Thalal, A. Modelling Electrostatic Potential from Experimentally Determined Charge Densities. II. Total Potential. *Acta Crystallogr., Sect. A* **1997**, *53*, 556–563.

[427] Yang, W. Direct Calculation of Electron Density in Density-Functional Theory. *Phys. Rev. Lett.* **1991**, *66*, 1438–1441.

[428] Galli, G.; Parrinello, M. Large Scale Electronic Structure Calculations. *Phys. Rev. Lett.* **1992**, *69*, 3547–3550.

[429] Goedecker, S. Linear Scaling Electronic Structure Methods. *Rev. Mod. Phys.* **1999**, *71*, 1085–1123.

[430] Ceriotti, M.; Kühne, T. D.; Parrinello, M. An Efficient and Accurate Decomposition of the Fermi Operator. *J. Chem. Phys.* **2008**, *129*, 24707.

[431] Fedorov, D. G.; Kitaura, K. Extending the Power of Quantum Chemistry to Large Systems with the Fragment Molecular Orbital Method. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.

[432] Merz, K. M. Using Quantum Mechanical Approaches to Study Biological Systems. *Acc. Chem. Res.* **2014**, *47*, 2804–2811.

[433] Walker, P. D.; Mezey, P. G. Molecular Electron Density Lego Approach to Molecule Building. *J. Am. Chem. Soc.* **1993**, *115*, 12423–12430.

[434] Meyer, B.; Guillot, B.; Ruiz-Lopez, M. F.; Genoni, A. Libraries of Extremely Localized Molecular Orbitals. 1. Model Molecules Approximation and Molecular Orbitals Transferability. *J. Chem. Theory Comput.* **2016**, *12*, 1052–1067.

[435] Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

[436] Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, Without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

[437] Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

[438] Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.

[439] Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. Machine Learning of Molecular Properties: Locality and Active Learning. *The Journal of Chemical Physics* **2018**, *148*, 241727.

[440] Hirshfeld, F. L. Bonded-Atom Fragments for Describing Molecular Charge Densities. *Theor. Chim. Acta* **1977**, *44*, 129–138.

[441] Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford Univer. Press, 1990.

[442] Gonthier, J. F.; Steinmann, S. N.; Wodrich, M. D.; Corminboeuf, C. Quantification of "fuzzy" Chemical Concepts: A Computational Perspective. *Chem. Soc. Rev.* **2012**, *41*, 4671.

[443] VandeVondele, J.; Hutter, J. Gaussian Basis Sets for Accurate Calculations on Molecular Systems in Gas and Condensed Phases. *J. Chem. Phys.* **2007**, *127*, 114105.

[444] Dominiak, P. M.; Volkov, A.; Dominiak, A. P.; Jarzembska, K. N.; Coppens, P. Combining Crystallographic Information and an Aspherical-Atom Data Bank in the Evaluation of the Electrostatic Interaction Energy in an Enzyme–substrate Complex: Influenza Neuraminidase Inhibition. *Acta Crystallogr., Sect. D* **2009**, *65*, 485–499.

[445] Pichon-Pesme, V.; Jelsch, C.; Guillot, B.; Lecomte, C. A Comparison Between Experimental and Theoretical Aspherical-Atom Scattering Factors for Charge-Density Refinement of Large Molecules. *Acta Crystallogr., Sect. A* **2004**, *60*, 204–208.

[446] Stewart, R. F. Electron Population Analysis with Generalized X-Ray Scattering Factors: Higher Multipoles. *J. Chem. Phys.* **1973**, *58*, 1668.

[447] Stewart, R. F. Electron Population Analysis with Rigid Pseudoatoms. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32*, 565–574.

[448] Hansen, N. K.; Coppens, P. Testing Aspherical Atom Refinements on Small-Molecule Data Sets. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *34*, 909–921.

[449] Coppens, P.; Guru Row, T. N.; Leung, P.; Stevens, E. D.; Becker, P. t.; Yang, Y. W. Net Atomic Charges and Molecular Dipole Moments from Spherical-Atom X-Ray Refinements, and the Relation Between Atomic Charge and Shape. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1979**, *35*, 63–72.

[450] Jelsch, C.; Pichon-Pesme, V.; Lecomte, C.; Aubry, A. Transferability of Multipole Charge-Density Parameters: Application to Very High Resolution Oligopeptide and Protein Structures. *Acta Crystallogr. D* **1998**, *54*, 1306–1318.

[451] Zarychta, B.; Pichon-Pesme, V.; Guillot, B.; Lecomte, C.; Jelsch, C. On the Application of an Experimental Multipolar Pseudo-Atom Library for Accurate Refinement of Small-Molecule and Protein Crystal Structures. *Acta Crystallogr. A* **2007**, *63*, 108–125.

[452] Lecomte, C.; Jelsch, C.; Guillot, B.; Fournier, B.; Lagoutte, A. Ultrahigh-Resolution Crystallography and Related Electron Density and Electrostatic Properties in Proteins. *J. Synchrotron Radiat.* **2008**, *15*, 202–203.

[453] Domagala, S.; Munshi, P.; Ahmed, M.; Guillot, B.; Jelsch, C. Structural Analysis and Multipole Modelling of Quercetin Monohydrate – a Quantitative and Comparative Study. *Acta Crystallogr. B* **2011**, *67*, 63–78.

[454] Domagala, S.; Fournier, B.; Liebschner, D.; Guillot, B.; Jelsch, C. An Improved Experimental Databank of Transferable Multipolar Atom Models – ELMAM2. Construction Details and Applications. *Acta Crystallogr. A* **2012**, *68*, 337–351.

[455] Koritsanszky, T.; Volkov, A.; Coppens, P. Aspherical-Atom Scattering Factors from Molecular Wave Functions. 1. Transferability and Conformation Dependence of Atomic Electron Densities of Peptides Within the Multipole Formalism. *Acta Crystallogr. A* **2002**, *58*, 464–472.

[456] Dittrich, B.; Koritsánszky, T.; Luger, P. A Simple Approach to Nonspherical Electron Densities by Using Invarioms. *Angew. Chem., Int. Ed.* **2004**, *43*, 2718–2721.

[457] Hathwar, V. R.; Thakur, T. S.; Row, T. N. G.; Desiraju, G. R. Transferability of Multipole Charge Density Parameters for Supramolecular Synthons: A New Tool for Quantitative Crystal Engineering. *Cryst. Growth Des.* **2011**, *11*, 616–623.

[458] Glielmo, A.; Sollich, P.; De Vita, A. *Phys. Rev. B* **2017**, *95*, 214302.

[459] Contreras-García, J.; Johnson, E. R.; Keinan, S.; Chaudret, R.; Piquemal, J.-P.; Beratan, D. N.; Yang, W. NCIPLOT: A Program for Plotting Noncovalent Interaction Regions. *J. Chem. Theory Comput.* **2011**, *7*, 625–632.

[460] Deringer, V. L.; Csányi, G. *Phys. Rev. B* **2017**, *95*, 094203.

[461] De, S.; Bartók, A. A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids Across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

[462] Bartók, A. A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R. J.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, e1701816.

[463] Musil, F.; De, S.; Yang, J.; Campbell, J. E. J.; Day, G. G. M.; Ceriotti, M. Machine Learning for the Structure-Energy-Property Landscapes of Molecular Crystals. *Chemical Science* **2018**, *9*, 1289–1300.

[464] Elstner, M.; Seifert, G. Density Functional Tight Binding. *Philos. Trans. Royal Soc. A* **2014**, *372*, 20120483.

[465] Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

[466] Kapil, V. et al. I-PI 2.0: A Universal Force Engine for Advanced Molecular Simulations. *Comput. Phys. Commun.* **2018**, *in press.*

[467] Petraglia, R.; Nicolaï, A.; Wodrich, M. M. D.; Ceriotti, M.; Corminboeuf, C. Beyond Static Structures: Putting Forth REMD As a Tool to Solve Problems in Computational Organic Chemistry. *J. Comp. Chem.* **2016**, *37*, 83–92.

[468] Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials. *J. Chem. Phys.* **2018**, *148*, 241730.

[469] Stone, A. *International Series of Monographs on Chemistry*; Oxford University Press, 2013.

[470] Buckingham, A. D.; Fowler, P. W.; Hutson, J. M. Theoretical Studies of Van Der Waals Molecules and Intermolecular Forces. *Chem. Rev.* **1988**, *88*, 963–988.

[471] Castleman, Jr., A.; Hobza, P. van Der Waais Molecules II: Introduction. *Chem. Rev.* **1994**, *94*, 1721–1722.

[472] Brutschy, B.; Hobza, P. van Der Waals Molecules III: Introduction. *Chem. Rev.* **2000**, *100*, 3861–3862.

[473] Hobza, P.; Ȓezáč, J. Introduction: Noncovalent Interactions. *Chem. Rev.* **2016**, *116*, 4911–4912.

[474] Pastorczak, E.; Corminboeuf, C. Perspective: Found in Translation: Quantum Chemical Tools for Grasping Non-Covalent Interactions. *J. Chem. Phys.* **2017**, *146*, 120901.

[475] Porezag, D.; Pederson, M. R. Infrared Intensities and Raman-Scattering Activities Within Density-Functional Theory. *Phys. Rev. B* **1996**, *54*, 7830–7836.

[476] Gilson, M. K.; Honig, B. H. Calculation of Electrostatic Potentials in an Enzyme Active Site. *Nature* **1987**, *330*, 84–86.

[477] Mecozzi, S.; West, A. P.; Dougherty, D. A. Cation-Pi Interactions in Aromatics of Biological and Medicinal Interest: Electrostatic Potential Surfaces As a Useful Qualitative Guide. *Proc. Nati. Acad. Sci. USA* **1996**, *93*, 10566–10571.

[478] Sagara, T.; Klassen, J.; Ganz, E. Computational Study of Hydrogen Binding by Metal-Organic Framework-5. *J. Chem. Phys.* **2004**, *121*, 12543.

[479] Cardamone, S.; Hughes, T. J.; Popelier, P. L. A. Multipolar Electrostatics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 10367.

[480] Polavarapu, P. L. Ab Initio Vibrational Raman and Raman Optical Activity Spectra. *J. Phys. Chem.* **1990**, *94*, 8106–8112.

[481] Hughes, J. L. P.; Sipe, J. E. Calculation of Second-Order Optical Response in Semiconductors. *Phys. Rev. B* **1996**, *53*, 10751–10763.

[482] Sipe, J. E.; Shkrebtii, A. I. Second-Order Optical Response in Semiconductors. *Phys. Rev. B* **2000**, *61*, 5337–5352.

[483] Sharma, S.; Ambrosch-Draxl, C. Second-Harmonic Optical Response from First Principles. *Phys. Scr.* **2004**, *T109*, 128.

[484] Masunov, A. E.; Tannu, A.; Dyakov, A. A.; Matveeva, A. D.; Freidzon, A. Y.; Odinokov, A. V.; Bagaturyants, A. A. First Principles Crystal Engineering of Nonlinear Optical Materials. I. Prototypical Case of Urea. *J. Chem. Phys.* **2017**, *146*, 244104.

[485] Koritsanszky, T. S.; Coppens, P. Chemical Applications of X-Ray Charge-Density Analysis. *Chem. Rev.* **2001**, *101*, 1583–1628.

[486] Lecomte, C.; Guillot, B.; Muzet, N.; Pichon-Pesme, V.; Jelsch, C. Ultra-High-Resolution X-Ray Structure of Proteins. *Cell. Mol. Life Sci.* **2004**, *61*, 774–782.

[487] Jayatilaka, D.; Dittrich, B. X-Ray Structure Refinement Using Aspherical Atomic Density Functions Obtained from Quantum-Mechanical Calculations. *Acta Crystallogr. A* **2008**, *64*, 383–393.

[488] Schnieders, M. J.; Fenn, T. D.; Pande, V. S.; Brunger, A. T. Polarizable Atomic Multipole X-Ray Refinement: Application to Peptide Crystals. *Acta Crystallogr. D* **2009**, *65*, 952–965.

[489] Brunger, A.; Adams, P. *Comprehensive Biophysics*; Elsevier, 2012; pp 105–115.

[490] Gatti, C.; Macchi, P. In *Modern Charge-Density Analysis*; Gatti, C., Macchi, P., Eds.; Springer Netherlands: Dordrecht, 2012.

[491] Bader, R. F. W. A Quantum Theory of Molecular Structure and Its Applications. *Chem. Rev.* **1991**, *91*, 893–928.

[492] Bader, R. In *Oxford University Press*; Matta, C. F., Boyd, R. J., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2007.

[493] de Silva, P.; Corminboeuf, C. Simultaneous Visualization of Covalent and Noncovalent Interactions Using Regions of Density Overlap. *J. Chem. Theory Comput.* **2014**, *10*, 3745–3756.

[494] Johnson, E. R.; Keinan, S.; Mori-Sánchez, P.; Contreras-García, J.; Cohen, A. J.; Yang, W. Revealing Noncovalent Interactions. *J. Am. Chem. Soc* **2010**, *132*, 6498–6506.

[495] Walker, P. D.; Mezey, P. G. Molecular Electron Density Lego Approach to Molecule Building. *J. Am. Ceram. Soc.* **1993**, *115*, 12423–12430.

[496] Walker, P. D.; Mezey, P. G. Ab Initio Quality Electron Densities for Proteins: A MEDLa Approach. *J. Am. Chem. Soc.* **1994**, *116*, 12022–12032.

[497] Exner, T. E.; Mezey, P. G. Ab Initio-Quality Electrostatic Potentials for Proteins: An Application of the ADMa Approach. *J. Phys. Chem. A* **2002**, *106*, 11791–11800.

[498] Exner, T. E.; Mezey, P. G. Ab Initio Quality Properties for Macromolecules Using the ADMa Approach. *J. Comput. Chem.* **2003**, *24*, 1980–1986.

[499] Szekeres, Z.; Exner, T.; Mezey, P. G. Fuzzy Fragment Selection Strategies, Basis Set Dependence and HF-DFT Comparisons in the Applications of the ADMa Method of Macromolecular Quantum Chemistry. *Int. J. Quantum Chem.* **2005**, *104*, 847–860.

[500] Stoll, H.; Wagenblast, G.; Preuβ, H. on the Use of Local Basis Sets for Localized Molecular Orbitals. *Theor. Chim. Acta* **1980**, *57*, 169–178.

[501] Meyer, B.; Guillot, B.; Ruiz-Lopez, M. F.; Genoni, A. Libraries of Extremely Localized Molecular Orbitals. 1. Model Molecules Approximation and Molecular Orbitals Transferability. *J. Chem. Theory Comput.* **2016**, *12*, 1052–1067.

[502] Meyer, B.; Guillot, B.; Ruiz-Lopez, M. F.; Jelsch, C.; Genoni, A. Libraries of Extremely Localized Molecular Orbitals. 2. Comparison with the Pseudoatoms Transferability. *J. Chem. Theory Comput.* **2016**, *12*, 1068–1081.

[503] Meyer, B.; Genoni, A. Libraries of Extremely Localized Molecular Orbitals. 3. Construction and Preliminary Assessment of the New Databanks. *J. Phys. Chem. A* **2018**, *122*, 8965–8981.

[504] Hirshfeld, F. L. Difference Densities by Least-Squares Refinement: Fumaramic Acid. *Acta Crystallogr. B* **1971**, *27*, 769–781.

[505] Stewart, R. F. Electron Population Analysis with Rigid Pseudoatoms. *Acta Crystallogr. A* **1976**, *32*, 565–574.

[506] Hansen, N. K.; Coppens, P. Testing Aspherical Atom Refinements on Small-Molecule Data Sets. *Acta Crystallogr. A* **1978**, *34*, 909–921.

[507] Bogojeski, M.; Brockherde, F.; Vogt-Maranto, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Efficient Prediction of 3D Electron Densities Using Machine Learning. *Preprint* **2018**, *arXiv:1811.06255*.

[508] Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chem. Eur. J.* **2012**, *18*, 9955–9964.

[509] Grisafi, A.; Wilkins, D. M.; Willatt, M. J.; Ceriotti, M. Atomic-Scale Representation and Statistical Learning of Tensorial Properties. *Preprint* **2019**, *arXiv:1904.01623*.

[510] Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. on the Applicability of LCAO-Xα Methods to Molecules Containing Transition Metal Atoms: The Nickel Atom and Nickel Hydride. *Int. J. Quantum Chem. Symp.* **1977**, *11*, 81–87.

[511] Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. Auxiliary Basis Sets to Approximate Coulomb Potentials. *Chem. Phys. Lett.* **1995**, *240*, 283–290.

[512] Nagy, A.; March, N. H. Ratio of Density Gradient to Electron Density As a Local Wavenumber to Characterize the Ground State of Spherical Atoms. *Mol. Phys.* **1997**, *90*, 271–276.

[513] Bohórquez, H. J.; Boyd, R. J. on the Local Representation of the Electronic Momentum Operator in Atomic Systems. *J. Chem. Phys.* **2008**, *129*, 024110.

[514] Nagy, Á.; Liu, S. Local Wave-Vector, Shannon and Fisher Information. *Phys. Lett. A* **2008**, *372*, 1654–1656.

[515] Murray, J. S.; Brinck, T.; Lane, P.; Paulsen, K.; Politzer, P. Statistically-Based Interaction Indices Derived from Molecular Surface Electrostatic Potentials: A General Interaction Properties Function (GIPF). *J. Mol. Struct.* **1994**, *307*, 55–64.

[516] Murray, J. S.; Politzer, P. Statistical Analysis of the Molecular Surface Electrostatic Potential: An Approach to Describing Noncovalent Interactions in Condensed Phases. *J. Mol. Struct.* **1998**, *425*, 107–114.

[517] Bootsma, A. N.; Doney, A. C.; Wheeler, S. Predicting the Strength of Stacking Interactions Between Heterocycles and Aromatic Amino Acid Side Chains. *chemrxiv.7628939.v4* **2019**,

[518] Bootsma, A. N.; Wheeler, S. Converting SMILES to Stacking Interaction Energies. *chemrxiv.8079890.v1* **2019**,

[519] Volkov, A.; Koritsanszky, T.; Coppens, P. Combination of the Exact Potential and Multipole Methods (EP/MM) for Evaluation of Intermolecular Electrostatic Interaction Energies with Pseudoatom Representation of Molecular Electron Densities. *Chem. Phys. Lett.* **2004**, *391*, 170–175.

[520] Bootsma, A. N.; Wheeler, S. E. Tuning Stacking Interactions Between Asp–Arg Salt Bridges and Heterocyclic Drug Fragments. *J. Chem. Inf. Model.* **2019**, *59*, 149–158.

[521] Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

[522] Mundt, M.; Kümmel, S. Derivative Discontinuities in Time-Dependent Density-Functional Theory. *Phys. Rev. Lett.* **2005**, *95*, 203004.

[523] Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Discontinuous Nature of the Exchange-Correlation Functional in Strongly Correlated Systems. *Phys. Rev. Lett.* **2009**, *102*, 066403.

[524] Mirtschink, A.; Seidl, M.; Gori-Giorgi, P. Derivative Discontinuity in the Strong-Interaction Limit of Density-Functional Theory. *Phys. Rev. Lett.* **2013**, *111*, 126402.

[525] Yang, W.; Cohen, A. J.; Mori-Sánchez, P. Derivative Discontinuity, Bandgap and Lowest Unoccupied Molecular Orbital in Density Functional Theory. *J. Chem. Phys.* **2012**, *136*, 204111.

[526] Mori-Sánchez, P.; Cohen, A. J. The Derivative Discontinuity of the Exchange–correlation Functional. *Phys. Chem. Chem. Phys.* **2014**, *16*, 14378–14387.

[527] Tozer, D. J. Relationship Between Long-Range Charge-Transfer Excitation Energy Error and Integer Discontinuity in Kohn–Sham Theory. *J. Chem. Phys.* **2003**, *119*, 12697–12699.

[528] Zhao, Q.; Ioannidis, E. I.; Kulik, H. J. Global and Local Curvature in Density Functional Theory. *The Journal of Chemical Physics* **2016**, *145*, 054109.

[529] Janak, J. F. Proof That $\frac{\partial E}{\partial N_i} = \epsilon_i$ in Density-Functional Theory. *Phys. Rev. B* **1978**, *18*, 7165–7168.

[530] Egger, D. A.; Weissman, S.; Refaely-Abramson, S.; Sharifzadeh, S.; Dauth, M.; Baer, R.; Kümmel, S.; Neaton, J. B.; Zojer, E.; Kronik, L. Outer-Valence Electron Spectra of Prototypical Aromatic Heterocycles from an Optimally Tuned Range-Separated Hybrid Functional. *J. Chem. Theory Comput.* **2014**, *10*, 1934–1952.

[531] Srebro, M.; Autschbach, J. Tuned Range-Separated Time-Dependent Density Functional Theory Applied to Optical Rotation. *J. Chem. Theory Comput.* **2012**, *8*, 245–256.

[532] Refaely-Abramson, S.; Sharifzadeh, S.; Govind, N.; Autschbach, J.; Neaton, J. B.; Baer, R.; Kronik, L. Quasiparticle Spectra from a Nonempirical Optimally Tuned Range-Separated Hybrid Density Functional. *Phys. Rev. Lett.* **2012**, *109*, 226405.

[533] Kronik, L.; Stein, T.; Refaely-Abramson, S.; Baer, R. Excitation Gaps of Finite-Sized Systems from Optimally Tuned Range-Separated Hybrid Functionals. *J. Chem. Theory Comput.* **2012**, *8*, 1515–1531.

[534] Sun, H.; Autschbach, J. Electronic Energy Gaps for π-Conjugated Oligomers and Polymers Calculated with Density Functional Theory. *J. Chem. Theory Comput.* **2014**, *10*, 1035–1047.

[535] Whittleton, S. R.; Sosa Vazquez, X. A.; Isborn, C. M.; Johnson, E. R. Density-Functional Errors in Ionization Potential with Increasing System Size. *J. Chem. Phys.* **2015**, *142*, 184106.

[536] Ioannidis, E. I.; Kulik, H. J. Towards Quantifying the Role of Exact Exchange in Predictions of Transition Metal Complex Properties. *J. Chem. Phys.* **2015**, *143*, 034104.

[537] Gani, T. Z. H.; Kulik, H. J. Where Does the Density Localize? Convergent Behavior for Global Hybrids, Range Separation, and DFT+U. *J. Chem. Theory Comput.* **2016**, *12*, 5931–5945.

[538] Kraisler, E.; Kronik, L. Fundamental Gaps with Approximate Density Functionals: The Derivative Discontinuity Revealed from Ensemble Considerations. *J. Chem. Phys.* **2014**, *140*, 18A540.

[539] Kraisler, E.; Kronik, L. Elimination of the Asymptotic Fractional Dissociation Problem in Kohn-Sham Density-Functional Theory Using the Ensemble-Generalization Approach. *Phys. Rev. A* **2015**, *91*, 032504.

[540] Kraisler, E.; Schmidt, T.; Kümmel, S.; Kronik, L. Effect of Ensemble Generalization on the Highest-Occupied Kohn-Sham Eigenvalue. *J. Chem. Phys.* **2015**, *143*, 104105.

[541] Cococcioni, M.; de Gironcoli, S. Linear Response Approach to the Calculation of the Effective Interaction Parameters in the LDA+U Method. *Phys. Rev. B* **2005**, *71*, 035105.

[542] Dabo, I.; Ferretti, A.; Poilvert, N.; Li, Y.; Marzari, N.; Cococcioni, M. Koopmans' Condition for Density-Functional Theory. *Phys. Rev. B* **2010**, *82*, 115121.

[543] Chai, J.-D.; Chen, P.-T. Restoration of the Derivative Discontinuity in Kohn-Sham Density Functional Theory: An Efficient Scheme for Energy Gap Correction. *Phys. Rev. Lett.* **2013**, *110*, 033002.

[544] Ferretti, A.; Dabo, I.; Cococcioni, M.; Marzari, N. Bridging Density-Functional and Many-Body Perturbation Theory: Orbital-Density Dependence in Electronic-Structure Functionals. *Phys. Rev. B* **2014**, *89*, 195134.

[545] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.

[546] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

[547] Dral, P. O.; von Lilienfeld, O. A.; Thiel, W. Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2015**, *11*, 2120–2125.

[548] Bereau, T.; DiStasio, R. A.; Tkatchenko, A.; von Lilienfeld, O. A. Non-Covalent Interactions Across Organic and Biological Subsets of Chemical Space: Physics-Based Potentials Parametrized from Machine Learning. *J. Chem. Phys.* **2018**, *148*, 241706.

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

[550] Ramakrishnan, R.; von Lilienfeld, O. A. *Machine Learning, Quantum Chemistry, and Chemical Space*; Wiley Online Library, 2017; pp 225–256.

[551] Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309–3313.

[552] Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, e1701816.

[553] Ceriotti, M. Unsupervised Machine Learning in Atomistic Simulations, Between Predictions and Understanding. *J. Chem. Phys.* **2019**, *150*, 150901.

[554] Vydrov, O. a.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. Importance of Short-Range Versus Long-Range Hartree-Fock Exchange for the Performance of Hybrid Density Functionals. *J. Chem. Phys.* **2006**, *125*, 074106.

[555] Vydrov, O. A.; Scuseria, G. E. Assessment of a Long-Range Corrected Hybrid Functional. *J. Chem. Phys.* **2006**, *125*, 234109.

[556] Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning. *J. Chem. Phys.* **2018**, *148*, 241717.

[557] 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.

[558] Tamblyn, I.; Refaely-Abramson, S.; Neaton, J. B.; Kronik, L. Simultaneous Determination of Structures, Vibrations, and Frontier Orbital Energies from a Self-Consistent Range-Separated Hybrid Functional. *J. Phys. Chem. Lett.* **2014**, *5*, 2734–2741.

[559] Huang, B.; Symonds, N. O.; Lilienfeld, O. A. v. Quantum Machine Learning in Chemistry and Materials. *Handbook of Materials Modeling* **2018**, 1–27.

[560] Fias, S.; Heidar-Zadeh, F.; Geerlings, P.; Ayers, P. W. Chemical Transferability of Functional Groups Follows from the Nearsightedness of Electronic Matter. *Proc. Natl. Acad. Sci.* **2017**, *114*, 11633–11638.

[561] Unke, O. T.; Meuwly, M. A Reactive, Scalable, and Transferable Model for Molecular Energies from a Neural Network Approach Based on Local Information. *J. Chem. Phys.* **2018**, *148*, 241708.

[562] Okumoto, K.; Shirota, Y. Development of High-Performance Blue-Violet-Emitting Organic Electroluminescent Devices. *Appl. Phys. Lett.* **2001**, *79*, 1231–1233.

[563] Deotare, P. B.; Chang, W.; Hontz, E.; Congreve, D. N.; Shi, L.; Reusswig, P. D.; Modtland, B.; Bahlke, M. E.; Lee, C. K.; Willard, A. P.; Bulović, V.; Van Voorhis, T.; Baldo, M. A. Nanoscale Transport of Charge-Transfer States in Organic Donor–acceptor Blends. *Nat. Mater.* **2015**, *14*, 1130–1134.

[564] Jou, J.-H.; Kumar, S.; Agrawal, A.; Li, T.-H.; Sahoo, S. Approaches for Fabricating High Efficiency Organic Light Emitting Diodes. *J. Mater. Chem. C* **2015**, *3*, 2974–3002.

[565] Shahnawaz, S.; Sudheendran Swayamprabha, S.; Nagar, M. R.; Yadav, R. A. K.; Gull, S.; Dubey, D. K.; Jou, J.-H. Hole-Transporting Materials for Organic Light-Emitting Diodes: An Overview. *J. Mater. Chem. C* **2019**, *7*, 7144–7158.

[566] Körzdörfer, T.; Sears, J. S.; Sutton, C.; Brédas, J.-L. Long-Range Corrected Hybrid Functionals for π-Conjugated Systems: Dependence of the Range-Separation Parameter on Conjugation Length. *J. Chem. Phys.* **2011**, *135*, 204107.

[567] Refaely-Abramson, S.; Baer, R.; Kronik, L. Fundamental and Excitation Gaps in Molecules of Relevance for Organic Photovoltaics from an Optimally Tuned Range-Separated Hybrid Functional. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2011**,

[568] Jacquemin, D.; Moore, B.; Planchat, A.; Adamo, C.; Autschbach, J. Performance of an Optimally Tuned Range-Separated Hybrid Functional for 0–0 Electronic Excitation Energies. *J. Chem. Theory Comput.* **2014**, *10*, 1677–1685.

[569] Chi, W.-J.; Sun, P.-P.; Li, Z.-S. How to Regulate Energy Levels and Hole Mobility of Spiro-Type Hole Transport Materials in Perovskite Solar Cells. *Phys. Chem. Chem. Phys.* **2016**, *18*, 27073–27077.

[570] Chi, W.-J.; Li, Q.-S.; Li, Z.-S. Exploring the Electrochemical Properties of Hole Transport Materials with Spiro-Cores for Efficient Perovskite Solar Cells from First-Principles. *Nanoscale* **2016**, *8*, 6146–6154.

[571] Ghosh, D.; Isayev, O.; Slipchenko, L. V.; Krylov, A. I. Effect of Solvation on the Vertical Ionization Energy of Thymine: From Microhydration to Bulk. *J. Phys. Chem. A* **2011**, *115*, 6028–6038.

[572] Kotadiya, N. B.; Mondal, A.; Xiong, S.; Blom, P. W. M.; Andrienko, D.; Wetzelaer, G. A. H. Rigorous Characterization and Predictive Modeling of Hole Transport in Amorphous Organic Semiconductors. *Adv. Electron. Mater.* **2018**, *4*, 1800366.

[573] Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

[574] Ruiz, E.; Salahub, D. R.; Vela, A. Defining the Domain of Density Functionals: Charge-Transfer Complexes. *J. Am. Chem. Soc.* **1995**, *117*, 1141–1142.

[575] Ruiz, E.; Salahub, D. R.; Vela, A. Charge-Transfer Complexes: Stringent Tests for Widely Used Density Functionals. *J. Phys. Chem.* **1996**, *100*, 12265–12276.

[576] Dewar, M. J. S.; Ford, G. P. Relationship Between Olefinic .pi. Complexes and Three-Membered Rings. *J. Am. Chem. Soc.* **1979**, *101*, 783–791.

[577] Sinha, D.; Mukhopadhyay, S.; Chaudhuri, R.; Mukherjee, D. The Eigenvalue-Independent Partitioning Technique in Fock Space: An Alternative Route to Open-Shell Coupled-Cluster Theory for Incomplete Model Spaces. *Chem. Phys. Lett.* **1989**, *154*, 544–549.

[578] Stanton, J. F.; Gauss, J. Analytic Energy Derivatives for Ionized States Described by the Equation-of-motion Coupled Cluster Method. *J. Chem. Phys.* **1994**, *101*, 8938–8944.

[579] Dutta, A. K.; Neese, F.; Izsák, R. Towards a Pair Natural Orbital Coupled Cluster Method for Excited States. *J. Chem. Phys.* **2016**, *145*, 034102.

[580] Dutta, A. K.; Nooijen, M.; Neese, F.; Izsák, R. Exploring the Accuracy of a Low Scaling Similarity Transformed Equation of Motion Method for Vertical Excitation Energies. *J. Chem. Theory Comput.* **2018**, *14*, 72–91.

[581] Christensen, A. S.; Faber, F. A.; Huang, B.; Bratholm, L.; Tkatchenko, A.; Muller, K.; von Lilienfeld, O. A. QML: A Python Toolkit for Quantum Machine Learning. 2017; https://github.com/qmlcode/qml{%}0A.

[582] Himanen, L.; Jäger, M. O.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

[583] Pedregosa, F. et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

[584] Becke, A. D.; Savin, A.; Stoll, H. Extension of the Local-Spin-Density Exchange-Correlation Approximation to Multiplet States. *Theor. Chim. Acta* **1995**, *91*, 147–156.

[585] Li Manni, G.; Carlson, R. K.; Luo, S.; Ma, D.; Olsen, J.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration Pair-Density Functional Theory. *J. Chem. Theory Comput.* **2014**, *10*, 3669–3680.

[586] Carlson, R. K.; Li Manni, G.; Sonnenberger, A. L.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration Pair-Density Functional Theory: Barrier Heights and Main Group and Transition Metal Energetics. *J. Chem. Theory Comput.* **2015**, *11*, 82–90.

[587] Ghosh, S.; Sonnenberger, A. L.; Hoyer, C. E.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration Pair-Density Functional Theory Outperforms Kohn–Sham Density Functional Theory and Multireference Perturbation Theory for Ground-State and Excited-State Charge Transfer. *J. Chem. Theory Comput.* **2015**, *11*, 3643–3649.

[588] Hoyer, C. E.; Ghosh, S.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration Pair-Density Functional Theory Is As Accurate As CASPT2 for Electronic Excitation. *J. Phys. Chem. Lett.* **2016**, *7*, 586–591.

[589] Gagliardi, L.; Truhlar, D. G.; Li Manni, G.; Carlson, R. K.; Hoyer, C. E.; Bao, J. L. Multiconfiguration Pair-Density Functional Theory: A New Way to Treat Strongly Correlated Systems. *Acc. Chem. Res.* **2017**, *50*, 66–73.

[590] Ghosh, S.; Cramer, C. J.; Truhlar, D. G.; Gagliardi, L. Generalized-Active-Space Pair-Density Functional Theory: An Efficient Method to Study Large, Strongly Correlated, Conjugated Systems. *Chem. Sci.* **2017**, *8*, 2741–2750.

[591] Bao, J. L.; Odoh, S. O.; Gagliardi, L.; Truhlar, D. G. Predicting Bond Dissociation Energies of Transition-Metal Compounds by Multiconfiguration Pair-Density Functional Theory and Second-Order Perturbation Theory Based on Correlated Participating Orbitals and Separated Pairs. *J. Chem. Theory Comput.* **2017**, *13*, 616–626.

[592] Sand, A. M.; Hoyer, C. E.; Sharkas, K.; Kidder, K. M.; Lindh, R.; Truhlar, D. G.; Gagliardi, L. Analytic Gradients for Complete Active Space Pair-Density Functional Theory. *J. Chem. Theory Comput.* **2018**, *14*, 126–138.

[593] Sharma, P.; Truhlar, D. G.; Gagliardi, L. Active Space Dependence in Multiconfiguration Pair-Density Functional Theory. *J. Chem. Theory Comput.* **2018**, *14*, 660–669.

[594] Carlson, R. K.; Truhlar, D. G.; Gagliardi, L. On-Top Pair Density As a Measure of Left–Right Correlation in Bond Breaking. *J. Phys. Chem. A* **2017**, *121*, 5540–5547.

[595] Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.

[596] Peverati, R.; Truhlar, D. G. M11-L: A Local Density Functional That Provides Improved Accuracy for Electronic Structure Calculations in Chemistry and Physics. *J. Phys. Chem. Lett.* **2012**, *3*, 117–124.

[597] Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Function. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[598] Peverati, R.; Truhlar, D. G. Improving the Accuracy of Hybrid Meta-GGA Density Functionals by Range Separation. *J. Phys. Chem. Lett.* **2011**, *2*, 2810–2817.

[599] De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids Across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

[600] Mallat, S.; Zhang, Z. Adaptive Time-Frequency Decomposition with Matching Pursuits. [1992] Proc. IEEE-SP Int. Symp. Time-Frequency Time-Scale Anal. 1992; pp 7–10.

[601] Rubinstein, R.; Zibulevsky, M.; Elad, M. Efficient Implementation of the K-SVD Algorithm Using Batch Orthogonal Matching Pursuit. 2008.

Alberto
Fabrizio

EPFL SB ISIC LCMD
CH-1015 Lausanne
✆ *+41 (77) 459 75 37*
✉ *alberto.fabrizio@epfl.ch*
Birth date: September 16, 1992

Education

2016-Present **Ph.D. student in Theoretical Chemistry**, EPFL, Lausanne, Switzerland.
Director: Prof. Dr. C. Corminboeuf.

2014-2016 **M.Sc. in Chemistry**, EPFL, Lausanne, Switzerland.
Orientation: Theoretical and Physical Chemistry.
Master Thesis: “Overcoming the drawbacks of electronic structure methods, in the theoretical investigation of organic electronic materials”.
Director: Prof. Dr. C. Corminboeuf.

2011-2014 **B.Sc. in Chemistry**, EPFL, Lausanne, Switzerland.

Scientific Presentations

2019 **Talk: “Learning (from) the electron density: transferability, conformational/chemical diversity”** *Swiss Chemical Society Fall Meeting*, Zürich (Switzerland).

2019 **Poster: “A transferable machine-learning model of the electron density”** *MARVEL Review and Retreat*, Lausanne (Switzerland).

2019 **Poster: “A transferable machine-learning model of the electron density”** *MQM Conference*, Heidelberg (Germany).

2018 **Poster: “Does London dispersion influence the properties of molecules in the excited states?”** *ESPA*, Toledo (Spain).

2018 **Poster: “Symmetry-adapted machine-learning of the ground-state electron density”** *MARVEL Review and Retreat*, Lausanne (Switzerland).

2017 **Poster: “Does London dispersion influence the properties of molecules in the excited states?”** *Swiss Chemical Society Fall Meeting*, Basel (Switzerland).

2017 **Poster: “Can molecular catalysts break linear scaling relationships?”** *MARVEL Review and Retreat*, Lausanne (Switzerland).

2017 **Poster: “Accurate Electronic Structure Description of the Oxygen Evolution Reaction”** *MARVEL Site Visit*, Lausanne (Switzerland).

2016 **Talk: “Balancing London dispersion and the delocalization error with DFT functionals.”** *Swiss Chemical Society Fall Meeting*, Zürich (Switzerland).

2016 **Talk: “Balancing London dispersion and the delocalization error with DFT functionals.”** *Summer School of the SPP 1807 on London dispersion interactions*, Bremen (Germany).

Awards & Grants

2018 **SCS Chemistry Travel Award**, *SCNAT and SCS.*

2019 **SCS Metrohm Prize for the best oral presentation**, *SCS-Metrohm.*

2019 **Teaching Excellence Award**, *EPFL.*

Publications:

19. Fabrizio, A.; Petraglia R.; Corminboeuf, C. Balancing DFT Interaction Energies in Charged Dimers Precursors to Organic Semiconductors. **2019**, *ChemRxiv. Preprint.* https://doi.org/10.26434/chemrxiv.11309480.v1.

18. Fabrizio, A.: Briling, K.; Grisafi, A.; Corminboeuf, C. Learning (from) the Electron Density: Transferability, Conformational and Chemical Diversity. *CHIMIA*, **2020**, *accepted*.

17. Fabregat, R.; Fabrizio, A.; Meyer, B.; Hollas, D.; Corminboeuf, C. Hamiltonian-Resevoir Replica Exchange and Machine Learning Potentials for Computational Organic Chemistry. *J. Chem. Theory. Comput.* **2020**, *accepted*.

16. Fabrizio, A.; Meyer, B.; Corminboeuf, C. Machine Learning Models of the Energy Curvature versus Particle Number for Optimal Tuning of Long-Range Corrected Functionals. *J. Chem. Phys.* **2020**, *accepted.*

15. Fabrizio, A.; Meyer, B.; Fabregat, R.; Corminboeuf, C. Quantum Chemistry Meets Machine Learning. *CHIMIA* **2019**, *73* (12), 983–989.

14. Fabrizio, A.; Grisafi, A.; Meyer, B.; Ceriotti, M.; Corminboeuf, C. Electron Density Learning of Non-Covalent Systems. *Chem. Sci.* **2019**, *10* (41), 9424–9432.

13. Palumbo, C. T.; Barluzzi, L.; Scopelliti, R.; Zivkovic, I.; Fabrizio, A.; Corminboeuf, C.; Mazzanti, M. Tuning the Structure, Reactivity and Magnetic Communication of Nitride-Bridged Uranium Complexes with the Ancillary Ligands. *Chem. Sci.* **2019**, *10* (38), 8840–8849.

12. Liu, J.; Mishra, S.; Pignedoli, C. A.; Passerone, D.; Urgel, J. I.; Fabrizio, A.; Lohr, T. G.; Ma, J.; Komber, H.; Baumgarten, M.; et al. Open-Shell Nonbenzenoid Nanographenes Containing Two Pairs of Pentagonal and Heptagonal Rings. *J. Am. Chem. Soc.* **2019**, *141* (30), 12011–12020.

11. Liu, Y.; Varava, P.; Fabrizio, A.; Eymann, L. Y. M.; Tskhovrebov, A. G.; Planes, O. M.; Solari, E.; Fadaei-Tirani, F.; Scopelliti, R.; Sienkiewicz, A.; et al. Synthesis of Aminyl Biradicals by Base-Induced Csp^3–Csp^3 Coupling of Cationic Azo Dyes. *Chem. Sci.* **2019**, *10* (22), 5719–5724.

10. Sawatlon, B.; Wodrich, M. D.; Meyer, B.; Fabrizio, A.; Corminboeuf, C. Data Mining the C−C Cross-Coupling Genome. *ChemCatChem* **2019**, *11* (16), 4096–4107.

9. Suleymanov, A. A.; Ruggi, A.; Planes, O. M.; Chauvin, A.; Scopelliti, R.; Fadaei Tirani, F.; Sienkiewicz, A.; Fabrizio, A.; Corminboeuf, C.; Severin, K. Highly Substituted Triazolines: Solid-State Emitters with Electrofluorochromic Behavior. *Chem. Eur. J.* **2019**, *25* (27), 6718– 6721.
8. Falcone, M.; Barluzzi, L.; Andrez, J.; Fadaei Tirani, F.; Zivkovic, I.; Fabrizio, A.; Corminboeuf, C.; Severin, K.; Mazzanti, M. The Role of Bridging Ligands in Dinitrogen Reduction and Functionalization by Uranium Multimetallic Complexes. *Nat. Chem.* **2019**, *11* (2), 154–160.
7. Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.* **2019**, *5* (1), 57–64.
6. Busch, M.; Fabrizio, A.; Luber, S.; Hutter, J.; Corminboeuf, C. Exploring the Limitation of Molecular Water Oxidation Catalysts. *J. Phys. Chem. C* **2018**, *122* (23), 12404–12412.
5. Fabrizio, A.; Corminboeuf, C. How Do London Dispersion Interactions Impact the Photochemical Processes of Molecular Switches? *J. Phys. Chem. Lett.* **2018**, *9* (3), 464–470.
4. Cretenoud, J.; Özen, B.; Schmaltz, T.; Görl, D.; Fabrizio, A.; Corminboeuf, C.; Fadaei Tirani, F.; Scopelliti, R.; Frauenrath, H. Synthesis and Characterization of Semiaromatic Polyamides Comprising Benzofurobenzofuran Repeating Units. *Polym. Chem.* **2017**, *8* (14), 2197–2209.
3. Prlj, A.; Fabrizio, A.; Corminboeuf, C. Rationalizing Fluorescence Quenching in Meso-BODIPY Dyes. *Phys. Chem. Chem. Phys.* **2016**, *18* (48), 32668–32672.
2. Fabrizio, A.; Rotzinger, F. P. Quantum Chemical Study of the Water Exchange Mechanism of the Americyl(VI) Aqua Ion. *Inorg. Chem.* **2016**, *55* (21), 11147–11152.
1. Prlj, A.; Curchod, B. F. E.; Fabrizio, A.; Floryan, L.; Corminboeuf, C. Qualitatively Incorrect Features in the TDDFT Spectrum of Thiophene-Based Compounds. *J. Phys. Chem. Lett.* **2015**, *6* (1), 13–21.

Blue titles highlight first author or co-first author papers.