Quantum mechanics-based methods for the refinement of crystal structures and the analysis of non-covalent interactions

Erna K Wieduwilt

THÈSE

présentée et soutenue publiquement le 26 octobre 2021
pour l'obtention du titre de

DOCTEUR DE l'UNIVERSITÉ DE LORRAINE

Mention : Chimie

par

Erna K. Wieduwilt

Quantum mechanics-based methods for the refinement of crystal structures and the analysis of non-covalent interactions

Membres du jury :

Directeurs de thèse :

Dr. Alessandro Genoni	Université de Lorraine, Metz
Dr. Manuel F. Ruiz-López	Université de Lorraine, Nancy

Président de jury :

Prof. Benoît Guillot	Université de Lorraine, Nancy

Rapporteurs :

Dr. Sophie Sacquin-Mora	Laboratoire de Biochimie Théorique, Paris
Prof. Ulf Ryde	Lunds universitet, Lund

Examinateurs :

Prof. Anna Krawczuk	Universität Göttingen, Göttingen
Prof. Benoît Guillot	Université de Lorraine, Nancy

Laboratoire de Physique et Chimie Théoriques (LPCT) - UMR CNRS 7019
Université de Lorraine, 1 Boulevard Arago, 57070 Metz, France

Für all diejenigen, die immer hinter mir standen,
ganz besonders meine Eltern, Karolina & Ralf Wieduwilt

All models are wrong
but some are useful.
$\sim$ George Box

Acknowledgment

First of all, I would like to thank my supervisor, **Dr. Alessandro Genoni**. *Grazie mille*, for really exceptional and intensive supervision, for giving me the opportunity to learn new things and for always being open to discussions and other points of view. Thank you also for helping me to get along in France. *Merci beaucoup* to my second supervisor **Dr. Manuel F. Ruiz-López** for helping with many small and large things.

Besides my supervisors, I would also like to thank all the other members of my jury, **Dr. Sophie Sacquin-Mora**, **Prof. Ulf Ryde**, **Prof. Anna Krawczuk** and **Prof. Benoît Guillot**, for reading and evaluating my thesis, for coming to Metz for my defense and for making it an unforgettable experience.

Grazie mille, **Dr. Giovanni Macetti**, for working together with me on many of the projects and for teaching me how to use the QM/ELMO technique. Thank you for helpful discussions, beers, and very nice excursions in France.

Obtaining the results that are collected in this thesis would never have been possible without the help of our collaborators. *Dankeschön* to **Prof. Dylan Jayatilaka**, **Dr. Simon Grabowsky**, **Dr. Florian Kleemiß** and **Dr. Lorraine A. Malaspina**. You taught me the basics of crystallography and quantum chemistry, and the beauty of research. Thank you for collaborating with us on most of the HAR projects. *Gracias* to **Dr. Julia Contreras-García**, **Dr. David Arias-Olivares**, **Dr. Francesca Peccati** and **Dr. Rubén Laplaza**, for working with us on the NCI projects and for visiting us in Metz and inviting me to Paris. *Merci* to **Prof. Eric Hénon**, **Prof. Jean-Charles Boisson** and **Prof. Giancarlo Terraneo** for collaborating with us on the IGM projects. *Grazie* to **Prof. Piero Macchi** and **Dr. Rebecca Scatena** for helping to kick off the extension of the ELMO libraries to MOFs. It has been a pleasure to work with all of you.

Merci to the people in the lab in Metz: **Prof. Lorenzo Ugo Ancarani**, **Prof. Jean-Christophe Tremblay**, **Dr. Arnauld Leclerc**, **Prof. Bedors Joulakian**, **Prof. Claude Dal Cappello**, **Philippe Senot**, **Dr. Abdallah Ammar** and **Ambre Blanc**. I enjoyed working in such an international environment. Thank you for the Friday mornings with brioche and for political discussions. *Merci beaucoup* to **Philippe Senot**, for always making *Witze*, for your patience with me and my French skills and for your help. *Merci beaucoup* also to **Fabien Pascale**, for your work on the cluster. The whole thesis would be pretty empty without it! *Merci beaucoup* also to **Prof. Gérald Monard** and **Dr. Nicolas Claiser** for very valuable advice. *Merci* also to my French teachers, especially **Anne Didelot**. The world needs more teachers like you.

I am very thankful for all the small and large encounters in the past three years. Thank you,

Julian Hille, **Karolin Maichle**, **Dr. Jingjing Shao**, **Gila Kopper**, **Joel Creutzberg**, **Ernst Dennis Larsson**, **Iria Bolano Losada**, **Asja Kröger**, **Max Davidson**, **Vishal Kumar Porwal**, **Valentin Anfray** and **Dr. Vedran Vuković** for all the nice moments we shared together in Metz, in Lund, or on different conferences. *Tack* to the Swedes for hosting me for the first three and a half months in the beginning of the pandemic. *Merci beaucoup* also to the couple that used to greet me every morning on my way to the university, although you will probably never read this, you made sure I had a smile on my face every morning.

There have been many other people who had a significant influence on my path. In this regard, I would like to thank my teachers at the Georg-Christoph-Lichtenberg-Schule in Kassel, Germany, in particular **Frau Mahlke-Harms**, **Herr Dr. Heuer**, **Frau Momberg** and **Herr Füller**, for encouraging me to think independently. I would also like to thank **Prof. Swiderek**, **Prof. Beckmann** and **Dr. Borrmann** from the university of Bremen for giving me a first taste in quantum chemistry and for their encouragement. Thank you, **Julie and Prof. Dylan Jayatilaka**, for welcoming me in Perth and for letting me into your world.

Finally, a special thank goes to all my friends and my family. It is not exaggerated to say that I would have never gotten this far without you.

Tausend Dank, **Denise Kraft**, **Anneke Dittmar**, **Manon Koch**, **Su-Min Choi**, **Faith Hundtoft** and **Florian Schönewald**, for being there although I am always busy and usually far away, for listening to my thoughts, my crazy ideas and my doubts. Thank you for helping whenever I get lost in translation and for showing me paths that I never thought of.

I would like to say *Dankeschön* to **Dr. Manfred Wieduwilt**, who showed me his impressive and ever growing mineral collection since I was a small child. *Vielen Dank* also to **Tante Lydia**, for preventing me from getting cold feet, and to **Christine Endres**, **Manfred Endres** and **Dr. Thomas Endres** for advice and new perspectives. *Evcharisto*, **Gabi and Dr. Ueli Schwarz**, for advice and guidance. A special *Dankeschön* goes to my brother, **Norbert Wieduwilt** who always believes in me. I am extremely grateful to my parents, **Karolina and Ralf Wieduwilt**. *Vielen, vielen Dank Mama und Papa* for endless support. From you I have the love for natural sciences and the deep belief that I am in the right place even when I am out of my depth sometimes. Thank you for roots and wings.

Last but not least, I would like to thank **Justin Bergmann**. Despite living in different countries or even on different continents most of the time, you were there for me, pretty much every day. I thank you for long discussions about work and life, and for even longer hiking tours. Thank you for feeding me while I was writing the thesis. Thank you for giving me shelter during the pandemic, thank you for staying with me and thank you for always having a compass when I lose my sense of direction.

Contents

Abbreviations

2D	two dimensional
3D	three dimensional
ADMA	adjustable density matrix assembler
ADP	anisotropic displacement parameter
AFDF	additive fuzzy density fragmentation
AO	atomic orbital
BODD	bond-oriented deformation density
CC	coupled cluster
CCSD	coupled cluster with single and double excitations
CCSD(T)	coupled cluster with single and double excitations and perturbative triples
CSD	Cambridge structural database
DC	"divide & conquer"
DFT	density functional theory
DFT/ELMO	QM/ELMO calculation at DFT level
DNA	deoxyribonucleic acid
E−H	element-hydrogen
EDA	energy decomposition analysis
ELF	electron localization function
ELI-D	electron localizability indicator
ELMO	extremely localized molecular orbital
ESP	electrostatic potential
FDET	frozen density embedding theory
FMO	fragment molecular orbital
GGA	generalised gradient approximation
HAR	Hirshfeld atom refinement
HAR-ELMO	HAR combined with ELMO libraries
HAR-QM/ELMO	HAR combined with QM/ELMO
HF	Hartree-Fock
HF/ELMO	QM/ELMO calculation at HF level
IAM	independent atom model
IGM	independent gradient model
IGM-ELMO	independent gradient model - extremely localized molecular orbital
IQA	interacting quantum atom
KEM	kernel energy method
KS	Kohn-Sham

LCAO	linear combination of atomic orbitals
LDA	local density approximation
LSCF	local self-consistent field
MD	molecular dynamics
MEDLA	molecular electron density lego assembler
MFCC	molecular fractionation with conjugate caps
MLDFT	mutilevel density functional theory
MLHF	mutilevel Hartree-Fock
MLCC	multilevel coupled cluster
MM	molecular mechanics
MO	molecular orbital
MOF	metal organic framework
MP2	second-order Møller-Plesset
MTA	molecular tailoring approach
NBO	natural bond orbital
NCI	non-covalent interaction
NCI-ELMO	non-covalent interaction - extremely localized molecular orbital
NMR	nuclear magnetic resonance
NPD	non-positive definite
ONIOM	our Own N-layer Integrated molecular Orbital molecular Mechanics
PbE	projection-based embedding
PDB	protein data bank
SCF	self-consistent field
RDG	reduced density gradient
SAPT	symmetry adapted perturbation theory
TAAM	transferable aspherical atom model
QTAIM	quantum theory of atoms in molecules
QM	quantum mechanics
QM/ELMO	quantum mechanics / extremely localized molecular orbital
QM/ELMO/MM	quantum mechanics / extremely localized molecular orbital / molecular mechanics
QM/MM	quantum mechanics / molecular mechanics
WF-in-DFT	wavefunction-in-DFT

Résumé

Le monde qui nous entoure, tout ce que nous pouvons voir, toucher, sentir ou goûter, ainsi que nous-mêmes, tout cela est constitué d'atomes et molécules. Ces constituants élémentaires de la matière réagissent et interagissent les uns avec les autres de diverses manières. Pour étudier ces phénomènes, la chimie offre une vaste gamme de techniques expérimentales et théoriques. Par exemple, les expériences de cristallographie aux rayons X sont couramment utilisées pour obtenir des structures tridimensionnelles fiables de systèmes chimiques. Cependant, pour bien comprendre les propriétés chimiques, les interactions et les réactions des atomes et des molécules, il faut non seulement connaître leur disposition géométrique, mais aussi leur structure électronique. Malheureusement, à l'exception des systèmes constitués par un seul électron, les équations permettant d'étudier la structure électronique des atomes et des molécules sont beaucoup trop compliquées pour être résolues de manière analytique. Pour cette raison, dans le domaine de la chimie quantique, des nombreuses techniques approximatives ont été développées et sont aujourd'hui appliquées de manière routinière.

Dans le cadre de cette thèse, les calculs de chimie quantique ont été utilisés pour accomplir principalement deux tâches : (i) l'obtention de structures cristallines aux rayons X plus précises, et (ii) l'analyse des interactions non covalentes. En particulier, trois techniques jouent un rôle majeur dans le travail présenté dans cette thèse. La première est la méthode « Hirshfeld atom refinement » (HAR), qui est une stratégie émergente pour le raffinement des structures à rayons X. Les deux autres techniques sont la méthode « non-covalent interaction » (NCI) et la méthode « independent gradient model » (IGM), qui sont utilisées pour l'analyse des interactions non covalentes. Nous donnerons plus de détails sur ces trois techniques ci-dessous. Notons simplement ici que toutes ces techniques nécessitent un calcul préalable de la densité électronique. Cependant, les calculs entièrement basés sur la mécanique quantique (QM) deviennent irréalisables pour des grands systèmes. Pour cette raison, l'applicabilité de la méthode Hirshfeld atom refinement (HAR) est limitée par la taille des molécules étudiées. Pour les grands systèmes, les méthodes non-covalent interaction (NCI) et independent gradient model (IGM) ont recours généralement à l'approximation de la densité pro-moléculaire, où la densité électronique de la molécule examinée est décrite comme une somme de densités atomiques indépendantes et sphériques. Ces densités pro-moléculaires manquent de précision et sont bien connues pour fournir des résultats biaisés. Pour surmonter toutes ces limitations, il est nécessaire de faire appel donc à d'autres approches qui permettent d'accéder rapidement à des densités électroniques fiables.

Compte tenu du grand intérêt que présentent les grandes molécules comme les protéines, plusieurs approches QM ont été développées pour étudier ces systèmes. Ces approches ont en commun de subdiviser la grande molécule étudiée en différentes sous-unités plus petites. Il y a deux façons principales de réaliser cette division. La première consiste à diviser le grand système en plusieurs petits fragments, pour lesquels la propriété d'intérêt est calculée individuellement et les résultats sont ensuite recombinés pour fournir une estimation de la

propriété du système complet. Une autre méthode consiste à diviser le système en une sous-unité d'intérêt particulier plus une région environnante. Dans ce cas, la sous-unité d'intérêt est traitée à un niveau de chimie quantique élevé, tandis que l'environnement est décrit à un niveau inférieur. Les stratégies qui adoptent la première méthode sont appelées approches de fragmentation, tandis que les techniques basées sur la deuxième méthode sont appelées méthodes de « embedding ».

Une technique de fragmentation particulière qui joue un rôle central dans cette thèse est basée sur l'utilisation des « extremely localized molecular orbitals » (ELMOs). Les ELMOs sont des orbitales moléculaires strictement localisées sur de petits fragments moléculaires. Grâce à cette localisation stricte, elles peuvent être calculées sur de petites molécules, stockées dans des bases de données et ensuite transférées sur de systèmes plus grands pour reconstruire leurs fonctions d'onde. Les premières bases de données de ce type sont les bibliothèques extremely localized molecular orbital (ELMO) qui contiennent toutes les ELMOs des unités élémentaires (atomes, liaisons et groupes fonctionnels) des vingt acides aminés naturels. Le programme associé à ces bibliothèques est le logiciel *ELMOdb* qui permet de transférer les ELMOs aux structures des systèmes étudiés. De cette manière, les fonctions d'onde et les densités électroniques sont rapidement obtenues et peuvent être utilisées par exemple dans les raffinements de structures cristallines, ou pour les analyses de liaisons chimiques. Cependant, dans certaines situations, il est peut-être souhaitable de décrire une certaine région du système (par exemple, le site actif d'une protéine) à un niveau de mécanique quantique plus élevé. Pour cette raison l'approche ELMO est également à la base de la technique de « embedding » QM/ELMO, où la région d'intérêt est traitée à un niveau QM plus élevé (par exemple en utilisant des méthodes Hartree-Fock (HF), « density functional theory » (DFT) ou post-HF) et la partie restante du système est décrite avec des ELMOs gelées.

Les approximations des méthodes de chimie quantique standard, leurs limites, ainsi que plusieurs exemples de techniques de fragmentation et de « embedding » sont décrits dans les Chapitre 1 de cette thèse. Des détails supplémentaires sur les méthodes ELMO et QM/ELMO sont donnés dans les Sections 1.3 et 1.5, respectivement. Les applications de ces techniques au raffinement de structures cristallines et à l'analyse des interactions non covalentes constituent les principaux sujets du travail présenté dans cette thèse. Ce manuscrit est donc divisée en deux parties, avec la Partie I se concentrant sur le raffinement de structures cristallines, et la Partie II sur l'analyse des interactions non covalentes. La suite de ce résumé sera également divisée de la même manière.

Raffinement de structures cristallines basé sur des méthodes de mécanique quantique

Comme mentionné ci-dessus, les structures tridimensionnelles de petites et grandes molécules peuvent être déterminées de façon routinière et fiable par des expériences de diffraction des rayons X sur monocristal. De cette manière, il est possible d'obtenir des positions très précises des atomes autres que l'hydrogène. Les positions des atomes d'hydrogène sont, au contraire, généralement systématiquement décalées vers leur partenaire de liaison. De ce fait, les longueurs de liaison élément-hydrogène (element-hydrogen (E−H)) dans les structures aux rayons X sont généralement trop courtes. La raison de ce défaut réside en partie dans la nature de

l'expérience, mais de manière plus importante dans les modèles théoriques qui sont à la base des stratégies de raffinement.

En fait, les rayons X sont diffusés par les électrons du cristal. L'expérience correspondante peut donc donner accès à la densité électronique. Dans la plupart des situations, les maxima de cette densité correspondent aux positions des noyaux, mais pas dans tous les cas. Par exemple, les atomes d'hydrogène ne possèdent qu'un seul électron, qui est impliqué dans la liaison chimique. Ainsi, le maximum de densité électronique associé est déplacé selon la liaison et la densité atomique de l'atome d'hydrogène est essentiellement asphérique. Cet aspect doit être pris en compte lors du raffinement par le modèle théorique sous-jacent. Il est important de noter, en effet, que toute stratégie de raffinement est toujours basée sur un modèle. En particulier, la grande majorité des raffinements est basée sur le modèle « independent atom model » (IAM), qui utilise les sommes de densités atomiques tabulées et sphériques pour fournir la densité pro-moléculaire du système. Cependant, comme mentionné ci-dessus, la densité atomique des atomes d'hydrogène dans les molécules n'est pas sphérique. Pour tenir compte de cette asphéricité, il faut envisager des modèles qui vont au-delà du modèle des atomes indépendants (independent atom model (IAM)). À cette fin, la technique HAR a été développée. Elle est basée sur le calcul direct de la densité électronique pour la molécule examinée en utilisant des calculs QM. Il a été démontré que la technique HAR peut fournir des longueurs de liaison E−H très proches des valeurs de référence neutroniques. Une introduction plus détaillée au raffinement de structures cristallines aux rayons X et aux neutrons, à la technique HAR et aux approches liées est donnée dans la Chapitre 2 de cette thèse.

Dans la Section 2.3.4, la procédure complète de la méthode HAR est expliquée. Elle consiste généralement dans la répétition itérative des trois étapes suivantes : (i) calcul de la densité électronique par de techniques QM, (ii) partitionnement de Hirshfeld de la densité électronique calculée et calcul ultérieur des facteurs de structure et (iii) raffinement par la méthode des moindres carrés. Le travail présenté dans cette thèse se concentre sur la première étape et sur les possibles moyens d'améliorer les résultats HAR en changeant le type de calcul QM.

En particulier, le premier raffinement de la structure d'une protéine avec la méthode HAR sera discuté dans le Chapitre 3. Pour accomplir cette tâche, les bibliothèques ELMO ont été utilisées pour obtenir la densité électronique, donnant naissance à la nouvelle technique HAR-ELMO. Cette approche a d'abord été validée en raffinant les structures de l'acide aminé *L*-alanine et du dipeptide glycine-*L*-alanine. Les résultats de la technique HAR-ELMO ont été comparés à ceux obtenus par les raffinements traditionnels HAR, IAM et de diffraction de neutrons. La validation a permis de conclure que la technique HAR-ELMO fournit des longueurs des liaisons E−H aussi précises que la méthode HAR traditionnelle mais à un coût de calcul considérablement réduit. Pour cette raison, l'approche HAR-ELMO nous a permis d'affiner les structures de deux polypeptides et d'une petite protéine. Néanmoins, d'autres améliorations sont nécessaires avant que la technique puisse devenir applicable de manière routinière. En effet, la résolution d'une structure protéique typique est trop faible pour les raffinements HAR ou HAR-ELMO. De plus, la possibilité de tenir compte du désordre dans les techniques HAR et HAR-ELMO n'existait pas au moment de l'étude. Ce dernier inconvénient n'a été surmonté que très récemment, comme cela sera décrit plus en détail dans le Chapitre 6.

Outre l'accélération des raffinements, il est également intéressant d'améliorer l'approche HAR, en particulier pour les raffinements de données à très haute résolution de petites molécules. Une possibilité est de tester différentes méthodes QM pour calculer les densités électroniques utilisées avec la technique HAR, et de comparer les paramètres obtenus entre eux et avec les résultats neutroniques de référence. Cette possibilité a été envisagée dans le travail présenté dans le Chapitre 4. En particulier, les données à haute résolution et à basse température de la *L*-alanine ont été raffinées avec HAR en utilisant six méthodes QM différentes (ELMOs, HF, BLYP, B3LYP, second-order Møller-Plesset (MP2) et coupled cluster with single and double excitations (CCSD)) en combinaison avec trois bases d'orbitales atomiques (def2-SVP, def2-TZVP, et def2-TZVPP). Les différences résultantes dans les longueurs de liaison E−H et dans les paramètres d'agitation thermique (» anisotropic displacement parameters », ADPs) des atomes d'hydrogène étaient significativement plus petites que les écarts types associés à ces quantités. Par conséquent, d'après les raffinements de la *L*-alanine, il semble que l'utilisation de méthodes post-HF n'améliore pas les résultats du raffinement ; néanmoins, la réalisation de tests supplémentaires avec des systèmes différents serait souhaitable.

Un défaut particulier des raffinements HAR ou HAR-ELMO traditionnels est que les longueurs des liaisons E−H polaires, qui sont impliquées dans les interactions intermoléculaires avec d'autres molécules dans l'environnement du cristal, sont systématiquement plus courtes que celles des références neutroniques (bien qu'elles soient clairement plus longues que celles résultant des raffinements IAM). La raison est que les calculs traditionnels de la fonction d'onde QM pour HAR sont effectués en « phase gazeuse », c'est à dire, la molécule est traitée comme si elle était dans le vide. Lorsqu'un groupe de charges et de dipôles est placé autour de l'unité QM pour simuler l'environnement cristallin dans le calcul de la fonction d'onde, il est possible de montrer que les distances des liaisons E−H impliquées dans des contacts intermoléculaires courts sont généralement améliorées de manière significative, même si elles restent systématiquement plus courtes que les valeurs de référence neutroniques. Dans le Chapitre 5, une procédure alternative basée sur la technique QM/ELMO est proposée. En particulier, l'unité centrale QM est décrite à l'aide de la fonctionnelle B3LYP, tandis que les molécules environnantes dans un rayon défini sont également prises en compte dans le calcul en les décrivant avec des ELMOs gelées. Nous avons utilisé cette nouvelle approche « ELMO-embedded HAR » , pour le raffinement de la structure du xylitol. Nous avons pu observer qu'elle fournit des longueurs de liaison O−H en accord optimal avec les valeurs de référence neutroniques, ou du moins, systématiquement plus proches de celles-ci que les résultats de raffinements HAR sans "embedding » ou avec des "embeddings » classiques constitués de groupes de charges et de dipôles.

Enfin, dans le Chapitre 6, nos derniers travaux pour la mise en œuvre de la technique HAR-ELMO dans *olex2* sont présentées. *olex2* est un logiciel standard pour le raffinement de petites molécules. Une interface de *olex2* permettant d'effectuer un raffinement HAR traditionnel a été récemment développée et permet de surmonter de nombreuses limitations de la méthode HAR liées au logiciel, telles que l'absence de raffinement du désordre. Nous avons ajouté une nouvelle option à cette interface qui permet d'effectuer des raffinements HAR-ELMO. Elle est actuellement testée pour des raffinements de structures désordonnées de molécules biologiques. En outre, nous envisageons de mettre en œuvre la possibilité d'ef-

fectuer des raffinements HAR-QM/ELMO dans le futur. Nous prévoyons d'exploiter cette nouvelle technique pour le raffinement de structures d'hydrures organométalliques. Le raffinement de ces composés est particulièrement difficile car l'atome d'hydrogène d'intérêt est lié directement à un élément-trace métallique et la grande densité de l'atome lourd peut masquer le petit pic de densité de l'hydrogène. L'étude de ces systèmes exige non seulement d'excellentes mesures expérimentales, mais également d'un traitement minutieux du métal dans le processus de raffinement. Dans ce but, nous prévoyons d'exploiter la technique de « embedding » QM/ELMO, le métal et l'atome d'hydrogène lié étant décrits au niveau post-HF et le reste du système par des ELMOs gelées. De cette façon, nous visons à améliorer la description de l'atome métallique, à obtenir une meilleure précision de la longueur de liaison E−H, et à clarifier davantage si les techniques post-HF sont nécessaires pour la technique HAR.

En conclusion, les travaux présentés dans la première partie de la thèse montrent que (i) peut devenir significativement moins coûteuse en calcul si elle est combinée avec les bibliothèques ELMO, rendant même possible le raffinement de structures de protéines ; (ii) les méthodes post-HF ne sont probablement pas nécessaires pour améliorer les raffinements HAR de structures cristallines de petites molécules organiques ; (iii) des longueurs de liaisons E−H polaires plus précises peuvent être obtenues à partir de HAR en utilisant un « embedding » ELMO ; (iv) des méthodes efficaces pour effectuer des raffinements HAR, même pour des systèmes complexes, pourraient devenir réalité dans un avenir proche.

Analyse des interactions non covalentes dans des polypeptides et des protéines

Les interactions non covalentes ont une influence non négligeable sur les structures des molécules en phase liquide et en phase solide. Par exemple, elles déterminent dans une grande mesure les structures et les fonctions des macromolécules biologiques. Par conséquent, une description complète et précise du réseau d'interactions non covalentes est essentielle pour mieux comprendre les fonctions des systèmes biologiques et la dynamique des processus biologiques. Pour cette raison, de nombreuses stratégies différentes sont disponibles pour étudier ces interactions. On peut les regrouper grossièrement en différents types de méthodes selon qu'elles se basent sur des critères géométriques, des énergies d'interaction, ou des indicateurs dans l'espace réel. Une vue d'ensemble de ces techniques est donnée dans le Chapitre 8.

Deux techniques particulières qui jouent un rôle majeur dans la deuxième partie de la thèse sont les approches NCI et IGM, qui visent à révéler, classifier et également quantifier les interactions non covalentes. Les deux techniques appartiennent aux indicateurs dans l'espace réel car elles sont basées sur l'analyse simultanée de la densité électronique et d'une quantité basée sur le gradient de la densité électronique. En particulier, la technique NCI repose sur l'analyse du gradient reduit de la densité, tandis que la technique IGM utilise le descripteur appelé δg qui est plus directement lié au gradient de la densité électronique. Lorsque le gradient réduit de la densité ou le descripteur δg sont représentés en fonction de la densité électronique, des pics caractéristiques peuvent être observés. Ils sont les signatures des interactions non covalentes. De plus, les isosurfaces du gradient réduit de la densité électronique ou du descripteur δg peuvent être utilisées pour mettre en évidence les interactions non covalentes dans l'espace three dimensional (3D). Des détails sur les techniques NCI et IGM sont

donnés dans les Sections 9.1 et 10.1, respectivement.

Comme mentionné ci-dessus, les calculs entièrement QM de la densité électronique sont irréalisables pour des grands systèmes. Pour cette raison, les techniques NCI et IGM traditionnelles reposent sur l'approximation de la densité pro-moléculaire, ce qui entraîne des différences non négligeables dans les résultats par rapport à ceux obtenus avec les calculs entièrement QM. Pour résoudre ce problème, nous avons couplé les techniques NCI et IGM avec les bibliothèques ELMO. Les approches NCI-ELMO et IGM-ELMO résultantes et leurs performances sont décrites respectivement dans les Chapitres 9 et 10.

Le Chapitre 9 est divisé en deux parties. La première partie est consacrée à la validation de la technique NCI-ELMO sur les polypeptides et les protéines. Pour des interactions non covalentes sélectionnées dans les polypeptides, il est démontré que la méthode NCI-ELMO peut fournir des résultats qui sont plus fiables que ceux de la méthode NCI pro-moléculaire et qui sont très proches de ceux obtenus à partir des calculs NCI-density functional theory (DFT). En plus, ces exemples montrent également que la technique NCI-ELMO est significativement plus rapide que la stratégie NCI-DFT, ce qui permet l'étude des interactions non covalentes dans les protéines. Ceci est démontré dans le Chapitre 9 pour une variété d'interactions non covalentes dans différentes structures de protéines. Cependant, les conclusions précédentes ne sont basées que sur des analyses qualitatives. Nos développements actuels pour une analyse quantitative NCI-ELMO constituent le sujet principal de la deuxième partie du Chapitre 9. Pour accomplir cette tâche, nous avons adapté une stratégie initialement proposée pour quantifier les interactions non covalentes avec l'approche NCI pro-moléculaire. Cette stratégie est basée sur le calcul des valeurs des intégrales NCI pour les interactions intermoléculaires. Pour calculer les intégrales NCI, une paramétrisation préliminaire est nécessaire, ce qui est effectué dans la deuxième partie du Chapitre 9 pour la technique NCI-ELMO. Des comparaisons avec les intégrales associées à la méthode NCI pro-moléculaire ont montré que la méthode NCI-ELMO offre une meilleure corrélation avec les énergies d'interaction de référence. Enfin, nous avons également combiné la technique QM/ELMO avec la méthode NCI pour étudier les interactions dans des complexes protéine-ligand dans le but d'estimer les forces des interactions établies par différents résidus des protéines étudiées avec les ligands.

Dans le Chapitre 10, la procédure et les résultats de la validation de la technique IGM-ELMO sont décrits. La méthode originale IGM inclut de nombreux aspects de l'indice NCI, mais le descripteur δg dans IGM permet une analyse plus sélective des interactions. Nous avons utilisé la technique IGM-ELMO pour détecter les interactions non covalentes dans des polypeptides et des protéines. Nos tests de validation montrent que la technique IGM-ELMO est capable de détecter de manière fiable les interactions non covalentes dans des polypeptides. Elle fournit des résultats très proches des résultats de référence IGM-DFT, alors que la technique IGM pro-moléculaire est moins fiable. De plus, les interactions dans des systèmes similaires peuvent être classées correctement en fonction de leur force en utilisant la valeur des intégrales IGM-ELMO. Enfin, les résultats de la méthode IGM-ELMO peuvent être obtenus à un coût de calcul considérablement réduit par rapport à ceux de la méthode IGM-DFT. De ce fait, la nouvelle méthode peut également être appliquée à l'étude des interactions non covalentes dans les protéines, pour lesquelles elle fournit des tendances complètement analogues à celles observées pour les polypeptides. Ainsi, la technique IGM-ELMO peut être utilisée

pour identifier, classifier et classer les interactions non covalentes dans des polypeptides et des protéines.

En conclusion, le travail présenté dans la deuxième partie de la thèse montre que les techniques NCI-ELMO et IGM-ELMO représentent des outils fiables pour identifier les différentes interactions non covalentes dans des grands systèmes. Plus encore, en utilisant les stratégies NCI-ELMO et IGM-ELMO il est possible d'obtenir des informations sur la force de ces interactions.

Introduction

1 Quantum chemistry and its application to large systems

In 1929, Paul Dirac wrote that *"the underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are [...] completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation."*[1] This famous[2,3] quote will be 100 years old by the end of this decade and fundamental progress has certainly been made in the meantime, both in developing approximate methods and also in explaining the properties of atomic and molecular systems. Nevertheless, Dirac's statement was probably too optimistic regarding the computational cost that is generally required for quantum chemical calculations. This is especially the case if we want to compute highly accurate results or study very large systems. In this regard, the second sentence of the quote by Dirac has not lost any of its relevance. In fact, many approximate methods are still being developed today, with the goal of extending the previous limits of accuracy and of large scale computations.

The focus of this introductory chapter lies particularly on the different strategies that have been devised for tackling the second problem and that allow the study of very large systems by means of quantum mechanics. In general, these techniques can be grouped into two categories: fragmentation and embedding approaches. In this chapter, the general features of these strategies will be reviewed and a few selected techniques will be presented. Furthermore, a detailed description will be given for the fragmentation and embedding strategies that are developed in our lab and that have been extensively used to obtain the results presented in this thesis. However, before these specialized approaches will be discussed, let us first come back to the statement of Dirac and the reasons why he has probably been too optimistic. This will be explained in the next section, where some of the standard methods of quantum chemistry and their scaling behavior will be described.

1.1 Standard methods of quantum chemistry and their limitations

One of the central equations in quantum chemistry is the time-independent Schrödinger equation:[4]

$$\hat{H}\Psi = E\Psi \tag{1.1}$$

where $\hat{H}$ is the Hamiltonian operator, Ψ the wavefunction and E the total energy of the system. The time-independent Schrödinger Equation (1.1) is usually solved by completely neglecting relativistic effects and, for molecular systems, by additionally applying the Born-Oppenheimer approximation,[5] which assumes that the electrons are moving in the field of fixed nuclei and that the motion of the nuclei and electrons can be separated. In this way, the Schrödinger equation can be solved separately for the nuclear and the electronic problem. For the latter, the full Hamiltonian is reduced to the electronic Hamiltonian, which is composed of three terms: one for the kinetic energy of the electrons, another for the coulomb attraction between the nuclei and the electrons, and a third one for the repulsion between the electrons.[6]

Despite the previously specified approximations, the Schrödinger equation cannot be solved exactly for many-electron systems. Therefore, further strategies have been developed that allow us to find approximate solutions to Equation (1.1). Before some of these strategies will be briefly reviewed below, some important aspects about the solutions to the Schrödinger equation should be mentioned. Equation (1.1) represents an eigenvalue problem with eigenfunctions that are wavefunctions, and eigenvalues that are energies. According to the postulates of quantum mechanics,[7] the wavefunction intrinsically contains all the information about a system under exam and, in principle, the wavefunction can be used to compute the expectation values of any physical observable. An observable that is central to this thesis is the electron density, which gives the probability of finding an electron at any point of the real space. It can be computed from the square of the global wavefunction using the following equation

$$\rho(\boldsymbol{r}) = N \sum_{\xi=\alpha,\beta} \int ... \int |\Psi(\boldsymbol{r}, \xi, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)|^2 d\boldsymbol{x}_2 ... d\boldsymbol{x}_N. \tag{1.2}$$

where the square modulus of the global wavefunction is integrated over all the electron coordinates $\boldsymbol{x}_i = (\boldsymbol{r}_i, \xi_i)$ except for the first electron, for which $|\Psi|^2$ is summed over the two possible values of the spin coordinate $\xi = (\alpha, \beta)$.[8] As we will see below, the electron density plays a fundamental role in density functional theory (DFT). Moreover, and very importantly in the context of this thesis, the electron density is crucial for the refinement of X-ray crystal structures and it can be analyzed with specialized techniques to study covalent and non-covalent interactions in small and large systems (see Chapters 2 and 8, respectively).

1.1.1 The Hartree-Fock method

Let us now come back to the problem of solving the time-independent non-relativistic Schrödinger equation within the Born-Oppenheimer approximation. The underlying approximation of quantum chemistry is to assume that the wavefunction is a single Slater determinant[9] constructed from one-electron functions φ_i, which are usually called molecular orbitals (MOs)[10] or molecular spin orbitals, if the spin of the electrons is considered. The basic technique of modern quantum chemistry, the Hartree-Fock (HF) method,[11–15] searches for those MOs that variationally minimize[1] the energy associated with the single Slater determinant. This is equivalent to solving the following set of equations, called the Hartree-Fock equations:

$$\hat{f}(\boldsymbol{r})\varphi_i(\boldsymbol{r}) = \epsilon_i\varphi_i(\boldsymbol{r}) \tag{1.3}$$

where $\hat{f}$ is the Fock operator, a one-electron operator that is expressed as follows:

$$\hat{f}(\boldsymbol{r}) = \hat{h}(\boldsymbol{r}) + \hat{v}^{HF}(\boldsymbol{r}) \tag{1.4}$$

In Equation (1.4), $\hat{h}$ is the one-electron core Hamiltonian operator and $\hat{v}^{HF}$ is the average potential experienced by one electron due to the presence of all the other electrons. By introducing the average potential the many-electron problem is substituted with an approximate one-electron problem. In contrast to Equation (1.1), Equation (1.3) represents a pseudo eigenvalue problem since the Fock operator depends on its own solutions because the Hartree-Fock potential $\hat{v}^{HF}$ depends on the molecular orbitals φ_i. Therefore, the Hartree-Fock equations are nonlinear and need to be solved iteratively. This is done exploiting the self-consistent field (SCF) procedure, where an initial guess is made for the MOs, which are subsequently used to calculate the Hartree-Fock potential. Then Equations (1.3) are solved, which provides a new set of MOs for the computation of a new potential.[6] Repeating these steps will eventually lead to convergence and to self-consistency of the one-electron effective potential.

Nevertheless, for molecular systems, also the Hartree-Fock equations are still too complicated to be solved analytically and exactly. Instead they are solved approximately, by introducing a set of basis functions that is used to expand the molecular orbitals as follows:

$$\varphi_i(\boldsymbol{r}) = \sum_{\mu=1}^{K} C_{\mu i}\chi_\mu(\boldsymbol{r}), \tag{1.5}$$

where the coefficients $C_{\mu i}$ expand the generic orbital $\varphi_i(\boldsymbol{r})$ in the known basis $\{\chi_\mu(\boldsymbol{r})\}$. Basis functions are similar to atomic orbitals (AOs) and sometimes they are also called that way. For practical applications, they are almost always Gaussian functions and are available internally in quantum chemical programs or collected in basis set databases.[16–18] Equation (1.5) represents the linear combination of atomic orbitals (LCAO) approximation. With its help the problem of calculating the MOs in the Hartree-Fock Equation (1.3) is reduced to the task of computing the expansion coefficients $C_{\mu i}$.[6] In other words, the Hartree-Fock problem is reduced to one that can finally be solved exploiting standard matrix techniques.[6] The corre-

[1] According to the variational principle, the energy of an approximate trial wavefunction is always higher than the exact ground state energy. Therefore, in the variational method, the parameters of a trial wavefunction are varied until the corresponding expectation value for the energy reaches a minimum.[6]

sponding matrix expression is the Roothan-Hall[14,15] equation, which assumes the following form:

$$\boldsymbol{FC} = \boldsymbol{SCE}, \tag{1.6}$$

where $\boldsymbol{F}$ is the matrix representation of the Fock operator in the space spanned by the chosen basis, $\boldsymbol{C}$ is the matrix of the coefficients that expand the MOs in the space of the basis $\{\chi_\mu(\boldsymbol{r})\}$, $\boldsymbol{S}$ is the matrix of the overlap integrals between the basis functions and $\boldsymbol{E}$ is a diagonal matrix of the orbital energies. The overlap matrix $\boldsymbol{S}$ is due to the non-orthogonality of the basis functions.[6] Since the Fock matrix $\boldsymbol{F}$ depends on the expansion coefficients, also Equation (1.6) is nonlinear and needs to be solved iteratively using the SCF procedure outlined above.

Because of the fundamental role that the electron density will play in this thesis, it is worth mentioning that, within the LCAO approximation, the electron density can be expressed as follows:

$$\rho(\boldsymbol{r}) = \sum_{\mu,\nu} \chi^*_\mu(\boldsymbol{r}) P_{\mu\nu} \chi_\nu(\boldsymbol{r}) \tag{1.7}$$

where $P_{\mu\nu}$ is the density matrix. In case of a single Slater determinant for a $2N$-electron closed-shell system, the density matrix assumes the following form:

$$P_{\mu\nu} = 2 \sum_{i=1}^{N} C^*_{\mu i} C_{\nu i} \tag{1.8}$$

1.1.2 Post-Hartree-Fock methods

As mentioned above, in the HF method, the electron-electron repulsion is treated in an average way. Consequently, the correlation between the electrons is partially neglected in the HF approach. By using a Slater determinant the motion of the electrons with parallel spin is correlated,[6] thus, the so-called exchange or Fermi correlation is accounted for in HF. However, the Coulomb correlation is neglected since the motion of electrons with opposite spin is not correlated within the HF approach.[6] Hence, the HF method is incapable of producing reliable results for the study of chemical reactions, where electron correlation plays a major role.[19]

Nevertheless, on the basis of the HF approach, more advanced techniques have been developed with the aim of including electron correlation in the calculations. Here, only two of them are mentioned, namely, second-order Møller-Plesset (MP2)[20,21] and coupled cluster (CC) theroy. A commonly applied CC approach is coupled cluster with single and double excitations (CCSD)[22–25], which was used together with MP2 to calculate results in this thesis. The corresponding calculations start from occupied and virtual MOs that are typically obtained from a HF calculation. Electron correlation is included in the computations by expressing the wavefunction as a linear combination of several Slater determinants that describe the excitations of electrons from occupied to virtual MOs. The coefficients associated with the Slater determinants that appear in the wavefunction expansion are determined in a perturbative way in MP2 or using an exponential *ansatz* in CCSD.

1.1.3 Density functional theory

An alternative to the wavefunction-based techniques, which also includes electron correlation, is represented by density functional theory (DFT).[26,27] In 1964, Hohenberg and Kohn proved[28] that the exact ground state energy of a system is uniquely determined by the electron density. In other words, the energy is a functional of the electron density. This is also known as the first Hohenberg-Kohn theorem, while, according to the second theorem, the ground state energy of the system has a minimum corresponding to the correct density. Therefore, the variational principle applies also to electron densities. The advantage of DFT is that the electron density depends only on the three spatial coordinates, whereas wavefunctions depend on the $3N$ spatial coordinates and N spin coordinates of N electrons. Nevertheless, the exact functional dependence of the energy on the electron density remains unknown.[19] Faced with this difficulty, in 1965, Kohn and Sham[29] further developed the ideas of Hohenberg and Kohn by introducing the concept of a fictitious system of non-interacting electrons that are moving in an external potential, which is sometimes called the Kohn-Sham (KS) potential $v_{KS}(\boldsymbol{r})$. The key observation of KS-DFT is that the density of this non-interacting system is exactly the same as the density of the real interacting system.[26] In the KS method, the density and the ground state energy are obtained by variationally minimizing over a single determinant wavefunction,[27] which is equivalent to solving a set of equations similar to the Hartree-Fock equations (compare Equation (1.3)), namely, the Kohn-Sham equations:

$$\hat{f}_{KS}(\boldsymbol{r})\varphi_i(\boldsymbol{r}) = \epsilon_i\varphi_i(\boldsymbol{r}) \tag{1.9}$$

where $\hat{f}_{KS}$ is the Kohn-Sham one electron operator that takes the following form:

$$\hat{f}_{KS}(\boldsymbol{r}) = -\frac{1}{2}\boldsymbol{\nabla}^2 + v_{KS}(\boldsymbol{r}) \tag{1.10}$$

where the first term is the kinetic-energy operator and the second the Kohn-Sham potential. The orbitals that satisfy Equation (1.9) are the Kohn-Sham orbitals, which can be used to compute the density with:

$$\rho(\boldsymbol{r}) = \sum_{i=1}^{N} |\varphi_i(\boldsymbol{r})|^2 \tag{1.11}$$

The Kohn-Sham potential depends on the density, which in turn depends on the Kohn-Sham orbitals. Hence, Equations (1.9) depend on their own solution. Therefore, like the Hartree-Fock equations, the Kohn-Sham equations are solved iteratively exploiting the SCF procedure.[26,27] In summary, the procedure devised by Kohn and Sham is similar to the one exploited in the HF method, but includes a major part of exchange and correlation effects.[29]

For the non-interacting system, Kohn and Sham showed[29] that the total energy functional can be decomposed as follows:

$$E[\rho] = T_s[\rho] + \int v_{ne}(\boldsymbol{r})\rho(\boldsymbol{r})d\boldsymbol{r} + E_H[\rho] + E_{xc}[\rho] \tag{1.12}$$

where $T_s[\rho]$ is the non-interacting kinetic-energy functional, $v_{ne}(\boldsymbol{r})$ the nuclei-electron interaction, $E_H[\rho]$ the Hartree energy functional and $E_{xc}[\rho]$ the exchange-correlation func-

tional. Exact expressions are known for the first three terms in Equation (1.12), and only the exchange-correlation functional needs to be approximated.[27] Despite the contribution of the exchange-correlation energy E_{xc} to the total energy is usually small, its role should not be underestimated, as pointed out by Kurth and Perdew, who described E_{xc} as "*the principal ingredient of the glue that binds atoms together to form molecules and solids*".[30] Given its importance, one of the major challenges in DFT has always been the development of more and more accurate exchange-correlation functionals.[26] The different approximations to this functional have been ranked by Perdew, who introduced the following concept of the Jacob's ladder:[26,27,31,32]

- The first rung of the ladder is represented by the local density approximation (LDA),[29] where the expression for the energy takes the following generic form[26,27]:

$$E_{xc}^{LDA} = \int f(\rho) d\boldsymbol{r} \tag{1.13}$$

- In the generalised gradient approximation (GGA),[33] the second rung, additionally to the density also its gradient is considered:

$$E_{xc}^{GGA} = \int f(\rho, \boldsymbol{\nabla}\rho) d\boldsymbol{r} \tag{1.14}$$

- For the third rung, in the meta-generalised gradient approximation, also the second derivative and/or the non-interacting positive kinetic energy density $\tau(\boldsymbol{r})$[27] are introduced:

$$E_{xc}^{meta-GGA} = \int f(\rho, \boldsymbol{\nabla}\rho, \boldsymbol{\nabla}^2\rho, \tau) d\boldsymbol{r} \tag{1.15}$$

- For the next two rungs, a fraction of the exact exchange energy is added. The difference between the fourth and fifth rung is that the exact exchange energy is computed from the occupied orbitals in the former case, while in the latter one also virtual orbitals are included.[32]

Thus, the higher the rung in the ladder, the more sophisticated is the underlying approximation, which is usually accompanied by higher accuracy and larger computational cost. The goal of climbing the latter is to reach the heaven of chemical accuracy, which typically means that the errors in the resulting energy differences are below 1 kcal/mol.[31,32]

1.1.4 The computational cost associated with the different methods

In general, to estimate the required cost for a quantum chemical calculation, it is important to know how the computational resources R depend on the system size N that is either the number of atoms (in molecular mechanics) or the number of orbitals (in quantum mechanics).[34] This dependence is expressed in the scaling law

$$R = A \cdot N^B, \tag{1.16}$$

where the prefactor A is dominant mostly for small system sizes, while the exponent B becomes more important for large systems.[34] In Figure 1.1 different scaling behaviors are schematically compared. For example, the gold standard method of quantum chemistry, coupled cluster with

single and double excitations and perturbative triples (CCSD(T)),[35] scales with N^7. The cheaper CCSD and MP2 techniques still scale with N^6 and N^5, respectively.[36] The respective scaling of HF and DFT methods is quartic (N^4) or cubic (N^3).[34] Because of these scaling behaviors, the application of post-HF methods is often limited to a few tens of atoms,[34] while traditional DFT approaches may be applied to systems of several hundreds atoms.[37,38]

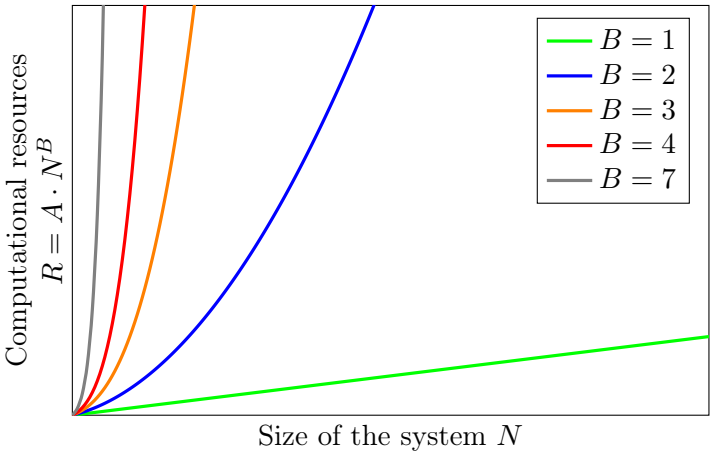


Figure 1.1: Schematic illustration of the scaling behavior of the computational resources depending on the size of the system.

However, due to the unfavorable scaling of the traditional methods in quantum chemistry, their application to very large systems (e.g. proteins and other biomolecules, polymers, solids) remains limited. For these systems, it is necessary to use different methods with a better scaling behavior. For example, quadratic (and also linear) scaling with the system size may be achieved with classical molecular mechanics (MM) force fields.[34] However, also in the field of quantum chemistry, the development of linearly scaling methods has been actively pursued. This was especially fueled in the 1990s, when the possibilities for parallel computations started to emerge.[34,39] At about the same time, different research groups devised some pioneering fragmentation techniques[40,41] that scale linearly with the system size. Today, several different fragmentation approaches[42–44] are available and some of them will be reviewed in the next section. Additionally to the fragmentation techniques, also embedding methods can significantly reduce the cost of computations on large systems and examples will be described in Section 1.4.

1.2 Fragmentation techniques

The general idea of the fragmentation techniques could be traced back to one of the four rules that the famous philosopher and scientist René Descartes (1596-1650) forced himself to strictly obey, namely *"to divide each of the difficulties I examined into as many parts as possible and as might be required in order to resolve them better."*[45] Also in the fragmentation techniques[42–44] of quantum chemistry, a large system is subdivided into several smaller, more manageable pieces. Individual quantum chemical calculations are subsequently performed on each of the fragments. Afterwards the obtained results are properly combined to give various properties of the entire system.

The aim of the fragmentation techniques is to approach the accuracy that a calculation on the complete system would have, but at a reduced computational cost. Because the computations on the subsystems are independent of each other, they can be performed in parallel. Linear (or nearly linear) scaling with the size of the system under exam is indeed achievable for fragmentation approaches, if the size of the fragments is independent of the target system size and if the number of fragments increases linearly with the size of the complete system.[43,44] Another advantage of the fragmentation approaches is that even computationally expensive quantum chemical techniques like post-HF methods may be applied for the calculations on the subunits, under the condition that the size of each individual subunit remains small enough.

The development of fragmentation techniques represents one of the most important tasks in the field of quantum chemistry.[44] Therefore, it is not surprising that a large number of fragmentation techniques[34,42–44] is available today. The database of extremely localized molecular orbitals (ELMOs), which was used to obtain almost all the results in this thesis, also belongs to this class of strategies. In the remaining part of this section, the key ideas for some related approaches will be summarized. Two of the pioneering fragmentation techniques, which were developed in the early 1990s, will be described at the beginning of this section. This will be followed by an introduction to fragment interaction methods and database approaches. The strategies described below represent only a small selection of the existing techniques (for other approaches see for example references [34, 42–44]).

1.2.1 Pioneering fragmentation techniques

The "divide & conquer" (DC) technique is one of the pioneering fragmentation techniques and was proposed by Yang[40,46] in 1991. It was one of the first linear scaling DFT methods and was developed in "the hope for [enabling] *ab initio* calculations of large systems beyond the reach of conventional methods"[46]. In the original DC method the electron density is the basic variable. The system under investigation is subdivided into different subsystems, e.g. atoms, functional groups or larger fragments. For each of the subsystems, the electron density is calculated separately, using a local basis set and a local Hamiltonian. Summing up the contributions from each of the subsystems, the total electron density and the total energy of the complete system are obtained. The original DC method was applied to several molecules, such as N_2[40], benzene, a tetrapeptide[46] and fullerenes with up to 1000 carbon atoms[47].

In further developments of the DC strategy, Yang and Lee introduced[48] an approach that is based on a partitioning of the density matrix instead of dividing the electron density. The new approach was found to be more efficient and is applicable also to semiempirical methods,[49,50] HF and post-HF techniques.[51–54] The semiemperical variant of the DC method was successfully applied to study several different polypeptides and proteins.[49,50,55–57] In particular, with the DC technique a calculation on a protein of more than 9000 atoms was feasible on a "typical workstation" already in 1996.[56]

To reduce the truncation errors, which occur because local basis sets are used, buffers were introduced in the calculations for each individual fragment.[46] The buffers consist of atoms in the neighborhood of the subsystem and their atomic orbitals are added to the local basis set of the subsystem. The buffers have been further improved by Merz and coworkers,[49,50]

who suggested to use a three layered scheme (see Figure 1.2). In particular, each subunit is embedded in an inner and outer buffer region. The core region, which corresponds to the subunit under exam, is thus described more accurately, while the inner buffer is less accurately described and the outer region exists only for insulation.[50] The buffer atoms are used only in the SCF calculations, but they are excluded from the computation of the subsystem density matrices.

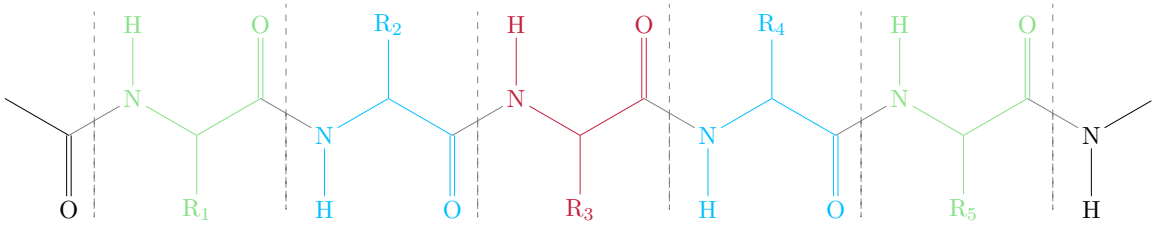


Figure 1.2: Definition of core (red) and inner (blue) and outer (green) buffer regions in a polypeptide.

The molecular tailoring approach (MTA)[41,58–62] is a method similar to the DC technique, which was developed independently by Garde and co-workers in 1994. The procedure of MTA consists of the following four steps: (i) Manual or automatic fragmentation of the complete system under exam into overlapping fragments. (ii) Set up of cardinality-based equations to compute the required quantities (e.g. the energy, the gradient, etc). In the framework of MTA, the cardinality[2] of a fragment refers to the number of atoms and bonds in the fragment. For example, the desired property P^M of a general molecule M that is subdivided into two fragments F_A and F_B, is given by the cardinality-based equation:

$$P^M = P^{F_A} + P^{F_B} - P^{F_A \cap F_B} \tag{1.17}$$

where P^{F_A} and P^{F_B} are the required properties of fragments F_A and F_B, respectively, and $P^{F_A \cap F_B}$ is the property for an overlapping fragment $(F_A \cap F_B)$ that is constituted of those atoms and bonds that are common to both the two fragments F_A and F_B (together with additional hydrogen atoms as caps). (iii) *Ab initio* calculations on all the fragments to obtain the required quantity for each fragment. For the general molecule M above, *ab initio* calculations are performed for fragments F_A, F_B and $F_A \cap F_B$. (iv) Recombination of the results for the individual fragments according to the cardinality-based equations to obtain the required quantities for the complete system.

To also include long-range interactions in the MTA, an electrostatically embedded molecular tailoring approach (EE-MTA)[63] was developed, where each of the fragments is embedded

[2] In mathematics, the cardinality refers to the number of elements in a set. For example, the cardinality $|A|$ of the set $A = \{1, 2, 3, 4, 5\}$ is five, while the cardinality of a second set $|B| = 3$ with $B = \{2, 4, 6\}$. According to the inclusion-exclusion principle, the cardinality of a set C, which contains the (unique) elements of set A and set B ($C = \{1, 2, 3, 4, 5, 6\}$), is given by:

$$\begin{aligned} |C| = |A \cup B| &= |A| + |B| - |A \cap B| \\ &= 5 + 3 - 2 = 6 \end{aligned}$$

in background charges, similar to the procedure used in quantum mechanics / molecular mechanics (QM/MM) (see Section 1.4.1). MTA can be applied to different types of systems. For example, peptides, proteins or nanotubes have been studied[64] as well as molecular crystal structures.[59] However, the MTA technique has been mostly applied to large molecular clusters[44,62], consisting for example of water or CO_2 molecules. In MTA, the individual and overlapping fragments are generally significantly larger[43,64] than in other fragmentation methods.

1.2.2 Fragment interaction methods

Similar to the molecular tailoring approach (MTA), in the fragment interaction techniques the properties of the system are computed from the sum of the properties of individual fragments and their intermolecular interactions. Therefore, the system is divided into individual fragments, which are afterwards completely or partially recombined to compute the contributions for the interactions. Three examples of fragment interaction techniques will be described below: the kernel energy method (KEM), the fragment molecular orbital (FMO) approach and the molecular fractionation with conjugate caps (MFCC) technique.

The kernel energy method (KEM)[65,66] is based on the division of the target system into individual kernels, which is done in such a way that each atom of the system is present in exactly one kernel. The dangling bonds are capped with hydrogen atoms. For each of the kernels, a quantum chemical calculation is performed. Additionally, the calculations are carried out for double kernels, which are all the possible pairwise combinations of two kernels. Also triple and quadruple kernels (for all possible combinations of three or four kernels, respectively) may be included to achieve higher accuracy.[67] The total energy for a system composed of n kernels is then calculated as

$$E_{\text{total}} = \sum_{1\leq i\leq n} E_i + \sum_{1\leq i<j\leq n} \Delta E_{ij} + \sum_{1\leq i<j<k\leq n} \Delta E_{ijk} + \sum_{1\leq i<j<k<l\leq n} \Delta E_{ijkl}, \tag{1.18}$$

where E_i is the energy of the single-kernel i, ΔE_{ij} is the interaction energy between the two kernels i and j, ΔE_{ijk} is the interaction energy between the three single kernels i, j and k and ΔE_{ijkl} is the interaction energy between the four single kernels i, j, k and l. The various interaction energies are given by:

$$\Delta E_{ij} = E_{ij} - (E_i + E_j) \tag{1.19}$$

$$\Delta E_{ijk} = E_{ijk} - (E_i + E_j + E_k) - (\Delta E_{ij} + \Delta E_{ik} + \Delta E_{jk}) \tag{1.20}$$

$$\begin{aligned}\Delta E_{ijkl} = E_{ijkl} &- (E_i + E_j + E_k + E_l) \\ &- (\Delta E_{ij} + \Delta E_{ik} + \Delta E_{il} + \Delta E_{jk} + \Delta E_{jl} + \Delta E_{kl}) \\ &- (\Delta E_{ijk} + \Delta E_{ijl} + \Delta E_{ikl} + \Delta E_{jkl})\end{aligned} \tag{1.21}$$

where E_{ij}, E_{ijl} and E_{ijkl} are the energies for the double, triple and quadruple kernels, respectively.

As mentioned above, all the possible combinations of single kernels to the higher order kernels are usually considered for the computation of the interaction energies. However, especially

in larger systems, many of the single kernels will be so far apart that their interaction energies will be approximately zero.[68] To avoid the calculation of these unnecessary interaction energies and to improve the computational efficiency of KEM, a more general model[68] has been developed, where only connected kernels are taken into account for building the higher order kernels. More precisely, the considered single kernels are connected through covalent bonds, including disulfide bridges in proteins, or through non-covalent interactions.[68]

KEM was applied to several biosystems including peptides[65] and proteins,[69–71] as well as DNA[72] and tRNA[73]. Furthermore, it was also possible to study a highly delocalized system such as graphene.[74]

The fragment molecular orbital (FMO) method[75–80] is nowadays a widely applied technique to study biosystems, nanomaterials and processes in solvation.[80] An FMO database has been recently assembled and currently includes more than 13 000 FMO calculations for biomolecules.[81] FMO computations may be based on various levels of theory,[79] and a multilayered FMO variant[82] is also available, where each of the layers may be described using different levels of theory and/or basis sets. Additionally, the FMO approach has been combined with tight binding DFT.[83] This strategy allows the study of very large systems, as was demonstrated with a geometry optimization of a fullerite cluster with more than 1 million atoms.[83]

The general FMO procedure starts with the fragmentation of the molecule, by assigning its electrons to different fragments. Usually single bonds are cut and the electrons forming these bond are attributed to one of the fragments. Unlike in many other fragmentation techniques, in FMO no hydrogen atoms or larger caps are added to the dangling bonds. In contrast, these bonds are saturated using an electrostatic field embedding (see below). The fragmentation procedure has been automatized for the systems usually studied within FMO, namely for polypeptides, saccharides and nucleotides.[78] A typical fragment contains ten to forty atoms.[76]

After the fragmentation, iterative monomer SCF calculations are performed. They consist of the following steps: (i) the electron density of each fragment is calculated without any embedding; (ii) the embedding for the different fragments is computed exploiting the densities obtained at the previous step; (iii) new electron densities are calculated with the electrostatic embedding computed at the preceding step. Steps (ii) and (iii) are repeated until convergence. In this way, the electron density obtained for each fragment is fully polarized by the electrostatic field due to all the other fragments. After the monomer SCF calculations are converged, further computations are performed on pairs of fragments with the goal of including also the charge transfer and the exchange-repulsion between the fragments. Furthermore, it is also possible to include triples and quadruples of fragments. Neither the pair nor the larger fragment computations are performed self-consistently. Instead the converged embedding obtained from the monomer SCF calculations is used also in the computations on the higher order fragments.

Finally, the total energy of the system is computed in the same way as in KEM, using Equations (1.18) to (1.21). Therefore, the main differences between FMO and KEM lie (i) in the capping for the KEM fragments but not for the FMO ones and (ii) in the electrostatic

embedding, which is applied in FMO but not in KEM.

The molecular fractionation with conjugate caps (MFCC) technique[84] was originally developed to compute the interaction energy between a protein P and an arbitrary molecule M. The underlying idea of the approach is that the interaction energy between the protein and the molecule is localized, so that the total interaction energy can be computed as the sum of the interactions between the molecule M and the individual fragments of the protein. The latter are usually obtained by cuts through the peptide bonds in the backbone (see the gray cutting indication in Figure 1.3). Conjugated caps are added to preserve the valency of the fragments (see capped fragments 1 and 2 in Figure 1.3). The conjugated caps are chosen in such a way that they resemble the local chemical environment. Additionally, so-called "concaps" are obtained by merging the two caps of neighboring fragments (see concap in Figure 1.3). The total interaction energy of the protein P with the molecule M is obtained by adding the individual interaction energies for each of the fragments with the molecule M and subtracting the contributions for the interactions between the molecule M and all the different concaps.

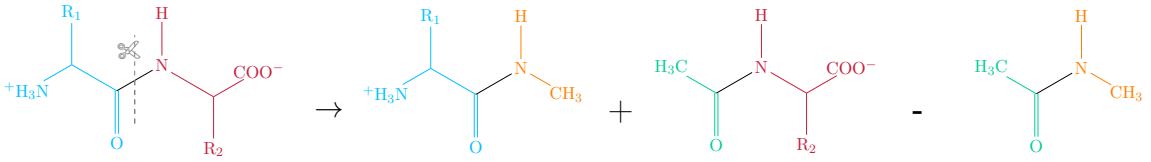


Figure 1.3: Example for a fragmentation with the MFCC approach: the peptide bond is cut and fragments 1 and 2 are capped using conjugated caps. From these fused caps, a concap fragment is formed.

Other than interaction energies, the MFCC strategy has been extended so that it is possible to compute the total electron density for proteins[85] and, from there, also the total energy by employing DFT. Further developments of the technique include a strategy based on density matrices, the MFCC-DM,[86] and the energy-corrected EC-MFCC,[87] which gives access to the ground state energies and geometries also for HF and post-HF methods. The MFCC technique has been recently coupled with the Hirshfeld atom refinement (HAR) strategy leading to the *fragHAR* approach.[88] *fragHAR* represents an alternative to the HAR-ELMO method, which will be presented in this thesis. Both approaches were developed with the goal of refining crystal structures of polypeptides and proteins using HAR (for more details, see Section 2.3.4 and Chapter 3).

1.2.3 Databanks of electron densities and density matrices

Contrary to the previously discussed approaches, where the quantum chemical calculations are performed directly on the individual fragments of the target system, in the database techniques pre-computed electron densities, density matrices or molecular orbitals are transferred from model molecules to the system under investigation. In the following paragraphs, two databases of electron densities and density matrices will be described. A library based on molecular

orbitals will be introduced in Section 1.3.

According to the additive fuzzy density fragmentation (AFDF) principle,[89–92] it is possible to divide the total electron density of a molecule into fuzzy overlapping fragment densities, which add up to the complete electron density of the molecule under exam:

$$\rho(\boldsymbol{r}, K) = \sum_{k=1}^{m} \rho^k(\boldsymbol{r}, K) \tag{1.22}$$

where $\rho^k(\boldsymbol{r}, K)$ is the additive fuzzy density of the k-th fragment at a point $\boldsymbol{r}$ for an explicit nuclear conformation K. The main realization of the AFDF principle was the development of two database techniques: the molecular electron density lego assembler (MEDLA)[93,94] and the adjustable density matrix assembler (ADMA)[95].

MEDLA[93,94] is a database of fuzzy electron density fragments, which were pre-calculated on small parent molecules and stored in cubic grid files. The electron density of a new molecule can then be obtained by translating and rotating the fragment densities in the database to the corresponding fragments in the new molecule. The MEDLA technique was applied to compute the electron densities of several proteins, for example crambin,[96] bovine insulin[97] and the HIV-1 protease monomer[98].

The MEDLA approach has the drawback that the electron density is stored in form of numerical grid files, which require a lot of memory.[91] This is avoided in the ADMA technique,[95] where the density matrices associated with the fragments are stored in a database. The accuracy obtained with ADMA is the same as for an ideal MEDLA variant, where numerical grids of infinite resolution would be used. Furthermore, with ADMA it is possible to compute a range of properties, such as the electron density, the electrostatic potential,[99,100] the dipole moment[100] as well as the Hartree-Fock energy[100] of proteins.

One of the main shortcomings of the ADMA approach is that the fragment density matrices do not include long range interactions.[34] A solution to partially overcome this problem is to compute the fragment density matrices from parent molecules that are directly derived from the target system and that are surrounded by point charges. This is called field-adapted ADMA approach,[101] which is significantly closer to the DC technique (see above) than the original databank-based ADMA method.[34]

1.3 Fragmentation based on extremely localized molecular orbitals

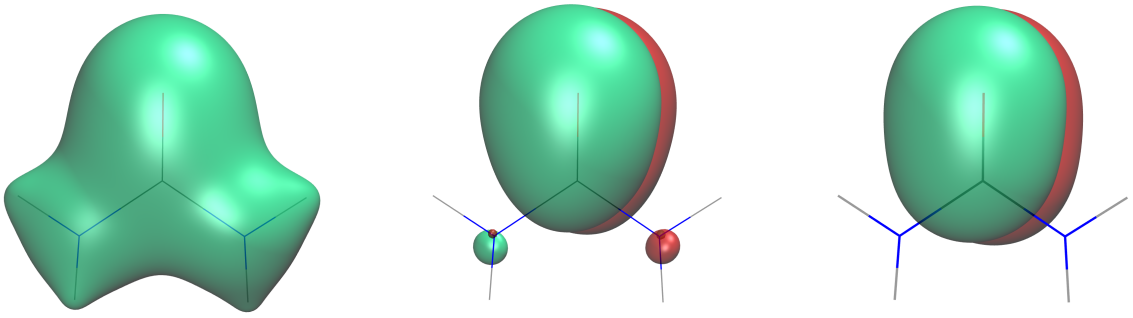
This section is dedicated to a detailed description of the fragmentation technique that plays a central role in the work presented in this thesis. Although it is related to the strategies described in the preceding section, it also differs in at least one important aspect. While the previously discussed fragmentation schemes were either energy-based or density-based, the technique described in this section is based on molecular orbitals (MOs) localized on fragments. In contrast, the method presented in this section also belongs to the database approaches like the previously introduced MEDLA and ADMA techniques.

However, not every kind of MO is suitable for a database, because the stored MOs must be transferable from a parent molecule to different target systems. To fulfill this condition, the

orbitals must be easily and unambiguously associable with small molecular units and, consequently, they must be strictly localized. In Section 1.3.1, standard techniques to localize MOs will be reviewed and their lacking suitability for a database approach will be discussed, while it will be shown that the so-called extremely localized molecular orbitals (ELMOs)[102–104] are ideal candidates for such a strategy. Hence, the procedures to compute and transfer them will be explained in Sections 1.3.2 and 1.3.3, respectively. Finally, the central fragmentation technique in this thesis, the ELMO libraries[105] will be introduced in Section 1.3.4.

1.3.1 Localization of molecular orbitals

Since the molecular orbitals obtained from standard MO-based quantum chemical calculations[3] are completely delocalized over the molecule (see Figure 1.4a), several approaches have been devised with the goal of computing localized MOs. Current methods may be grouped into *a posteriori* techniques, which transform the canoncial Hartree-Fock or Kohn-Sham MOs after the energy minimization, and *a priori* techniques, which constrain the shape of the MOs before the minimization of the energy.[107]



(a) Hartree-Fock (b) Pipek-Mezey (c) ELMO

Figure 1.4: Examples for a canonical Hartree-Fock orbital (a), for a Pipek-Mezey orbital (b) and for an extremely localized molecular orbital (c). All orbitals were calculated on the urea molecule with the cc-pVDZ basis set and were plotted using the 0.025 a.u. isosurface. While the Hartree-Fock orbital is completely delocalized over the entire molecule, the Pipek-Mezey orbital is mainly localized on the C−O bond except for the orthogonalization tails on the nitrogen atoms. In contrast, the ELMO is strictly localized on the C−O bond.

The strategies in the former group are based on unitary transformations of the canonical MOs that form a Slater determinant. Since determinants are invariant to unitary transformations, one set of MOs can be unitary transformed into a different set that still yields the same Slater determinant up to a phase factor and, consequently, the same minimized energy. There is no unique way for localizing the MOs using unitary transformations.[107] Therefore, several techniques have been developed. Examples include (i) the Foster-Boys method,[108,109] where the spatial extension of the MOs is minimized; (ii) the Edminston-Ruedenberg approach,[110,111] which aims at maximizing the sum of the self-repulsion energies of the orbitals; and (iii) the Pipek-Mezey technique,[112,113] where the charges associated with

[3] Alongside with the molecular orbital theory, also valence bond theory has been developed. Unlike the MOs, the valence bond orbitals are strictly localized and correspond closely to the Lewis picture.[106] However, the valence bond methods are computationally more expensive than the MO approaches and are usually applied to investigate quite small molecules. Therefore, valence bond techniques will not be further discussed in this thesis.

each orbital are maximized. Although all of the previously mentioned *a posteriori* techniques lead to MOs that are mainly localized on small molecular subunits, the obtained MOs are characterized by so-called "orthogonalization tails", which are necessary to maintain the orthogonality of the MOs. These tails extend to other parts of the molecule (see Figure 1.4b).[114] Therefore, MOs obtained from *a posteriori* techniques are not unambiguously associable with small molecular units and thus not transferable from one system to another, unless the tails are truncated, which leads to non-negligible errors in the energy.[103,115–117]

As mentioned above, MOs need to be strictly localized on small molecular fragments to be transferable. Such orbitals can be obtained using *a priori* techniques, where the MOs are localized before the energy minimization by introducing local basis sets. Thus, the number of variational parameters is reduced leading to a final wavefunction that is higher in energy than the corresponding single Slater determinant wavefunction constructed with canonical MOs.[107] Nevertheless, the associated error is still smaller than the one resulting from a truncation of the orthogonalization tails from orbitals computed with an *a posteriori* technique.[103] An example for an *a priori* technique is a method proposed by Stoll[102], where extremely localized molecular orbitals (ELMOs)[102–104] are calculated in a procedure similar to the HF one. The MOs obtained with this strategy are indeed strictly localized on small molecular fragments (see Figure 1.4c). Given the central role played by these orbitals in the work presented in this thesis, in the next subsection the Stoll technique will be presented with more details.

1.3.2 Computation of ELMOs

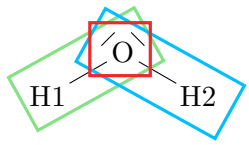
The procedure for the computation of ELMOs has been introduced in 1980 by Stoll and coworkers[102] and was rigorously implemented in a modified version[103] of the GAMESS-UK program[118]. In the first step of the procedure, a closed-shell molecule with $2N$ electrons is divided into subunits. In principle, this partitioning is arbitrary, although in most cases the molecule is divided according to its Lewis structure. Hence, three types of subunits are usually obtained: one-atom fragments for cores and lone pairs, two-atom fragments for bonds and three-atom fragments associated with functional groups. In principle, also larger fragments are possible, however, they will not be considered here.

The procedure of the ELMO computation will be explained on the example of the water molecule. According to its Lewis structure, the water molecule is divided into three fragments: a single one-atom fragment for the oxygen atom and two two-atom fragment (one for each O−H bond). The corresponding localization scheme is shown in Figure 1.5a.

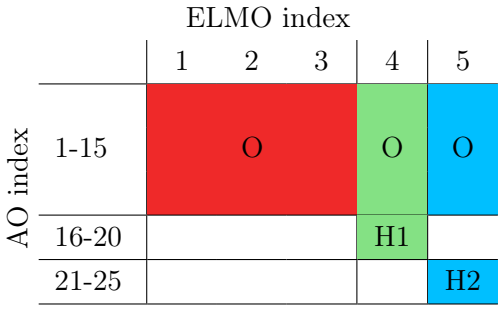
As briefly mentioned above, to compute ELMOs, a local basis set $\{\chi_{i\mu}\}_{\mu=1}^{M_i}$ is defined for each fragment, where M_i is the number of atomic orbitals that are centered only on the nuclei belonging to the i-th fragment. Thus, the α-th ELMO of the i-th fragment is given by:

$$|\varphi_{i\alpha}\rangle = \sum_{\mu=1}^{M_i} C_{i\mu,i\alpha} \; |\chi_{i\mu}\rangle \,, \tag{1.23}$$

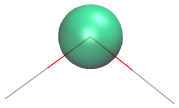
In this way, each ELMO is strictly localized only on the atoms of the corresponding fragment. For the example of the water molecule, the corresponding block-structured matrix of the



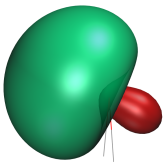
(a) Localization scheme



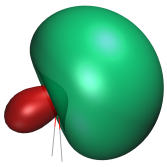
(b) Coefficient matrix



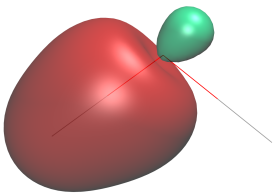
(c) ELMO 1



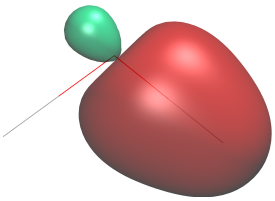
(d) ELMO 2



(e) ELMO 3



(f) ELMO 4



(g) ELMO 5

Figure 1.5: Localization scheme for the water molecule corresponding to its Lewis structure (a). Block-structured ELMO coefficient matrix for the cc-pVDZ basis set (b). Using the localization scheme in panel A, the following extremely localized molecular orbitals (ELMOs) for water were obtained: ELMO associated with the oxygen core (c), ELMOs corresponding to the oxygen lone pairs ((d) and (e)), ELMOs describing the O−H bonds ((f) and (g)). All orbitals were calculated with the cc-pVDZ basis set and were plotted using the 0.2 a.u. isosurface.

ELMO coefficients (for the basis set cc-pVDZ) is shown in Figure 1.5b. As can be seen in that figure, all five ELMOs share the basis functions that are centered on the oxygen atom (AO indices 1-15). In addition, the ELMO for the O−H1 bond is also expended on the basis functions that are located on atom H1 (AO indices 16-20), while for the O−H2 bond ELMO also the basis functions centered on atom H2 are used (AO indices 21-25). As can be seen already from the example of the water molecule, the constructed fragments overlap and share some basis functions. Hence, the ELMOs within one fragment are orthogonal to each other, but ELMOs between different fragments are not.

The ELMO wavefunction for the complete system is given by a single Slater determinant, which is composed of all the ELMOs in the system:

$$|\Psi_{\mathrm{ELMO}}\rangle = \frac{1}{\sqrt{(2N)!det[\tilde{\boldsymbol{S}}]}}\hat{A}\left[\prod_{i=1}^{f}\prod_{\alpha=1}^{n_i}\varphi_{i\alpha}\overline{\varphi_{i\alpha}}\right] \tag{1.24}$$

where $det[\tilde{\boldsymbol{S}}]$ is the determinant of the ELMO overlap matrix $\tilde{\boldsymbol{S}}$ due to the non-orthogonality of the ELMOs, $\hat{A}$ is the antisymmetrizer, which generates the Slater determinant, f is the total number of fragments and n_i is the number of occupied ELMOs for the i-th fragment. $\psi_{i\alpha}$ and $\overline{\psi_{i\alpha}}$ are spin orbitals, where the bar indicates the spin part β.

In a procedure analogous to the HF method, the ELMOs are obtained by variationally

minimizing the energy associated with the ELMO wavefunction (given by Equation (1.24)):

$$E[\varphi] = \langle \Psi_{\mathrm{ELMO}} | \hat{H} | \Psi_{\mathrm{ELMO}} \rangle \tag{1.25}$$

where Ψ_{ELMO} is the previously defined ELMO wavefunction, and $\hat{H}$ is the Hamiltonian operator. This is equivalent to solving the Stoll equations for each fragment:

$$\hat{F}_i \left| \varphi_{i\alpha} \right\rangle = \epsilon_i \left| \varphi_{i\alpha} \right\rangle \tag{1.26}$$

The Stoll equations are modified Hartree-Fock equations and include a modified Fock operator $\hat{F}_i$. For the i-th fragment, this operator is given by:

$$\hat{F}_i = (1 - \hat{\rho} + \hat{\rho}_i^{\dagger}) \hat{F} (1 - \hat{\rho} + \hat{\rho}_i) \tag{1.27}$$

where $\hat{F}$ is the usual Fock operator (used in the Hartree-Fock method) and $\hat{\rho}$ is the global density operator, which depends on all the occupied ELMOs in the system. Finally, $\hat{\rho}_i$ is the density operator, which is associated with fragment i and depends only on the occupied ELMOs of that fragment.

After introducing the previously defined local basis set $\{\chi_{i\mu}\}_{\mu=1}^{M_i}$ (see above) for fragment i, modified Roothan-Hall equations are obtained for that fragment:

$$\boldsymbol{F}_i \boldsymbol{C}_i = \boldsymbol{S}_i \boldsymbol{C}_i \boldsymbol{E}_i, \tag{1.28}$$

where the Fock matrix $\boldsymbol{F}_i$ is the matrix representation of the modified Fock operator given by Equation (1.27); $\boldsymbol{C}_i$ is the matrix of the coefficients for the ELMOs belonging to the i-th fragment; the overlap matrix $\boldsymbol{S}_i$ denotes the overlap between the atomic orbitals belonging to fragment i; and the diagonal matrix $\boldsymbol{E}_i$ contains the orbital energies for the ELMOs associated with fragment i. Equation (1.28) can be solved self-consistently for each subunit. However, the equations are coupled together because the modified Fock operator depends on the global density operator $\hat{\rho}$, which in turn depends on all the occupied ELMOs of the system. More details about the computation of ELMOs can be found in references [102, 103, 119].

Coming back to the example of the water molecule, the corresponding final ELMOs are shown in Figure 1.5. Each of the obtained ELMOs can be associated unambiguously with an electron pair in the Lewis structure. The fragment for the oxygen atom is associated with three individual ELMOs, namely one for the core (Figure 1.5c) and two for the lone pairs of the oxygen atom (Figures 1.5d and 1.5e). In contrast, the fragments for the O−H bonds include only one ELMO each, i.e. the one localized on the respective O−H bond (Figures 1.5f and 1.5g).

As mentioned above, the outlined procedure represents a modification of the HF strategy. However, ELMOs may also be computed in the framework of DFT[120] or obtained from fitting to X-ray structure factors in an approach[121] that adopts the X-ray constrained wavefunction fitting[122–129] originally developed by Jayatilaka and coworkers. Henceforth, whenever the term ELMO is used in this thesis, it refers to ELMOs that were computed with the modified HF strategy explained above. Nevertheless, it is important to note that ELMOs are not to be confused with HF orbitals, as was hopefully made clear in Figure 1.4.

1.3.3 Transfer and rotation of ELMOs

Since the computed ELMOs are strictly localized on small molecular units, they can be unambiguously associated with these units and are thus transferable from one molecule to another. The procedure[119] to transfer the ELMOs between different molecules is based on a strategy, which was developed by Philipp and Friesner[130] to transfer localized orbitals to the frontier bonds in QM/MM calculations (see Section 1.4.1). The key step of the procedure is the definition of a matrix $\boldsymbol{P}$ for the proper rotation of the ELMOs from the geometry in the model molecule to the one in the target molecule.

Before the rotation matrix $\boldsymbol{P}$ can be constructed, it is necessary to define two triads of atomic positions for each ELMO that needs to be transferred. The first triad $(\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3)$ is defined by the positions of three atoms in the model molecule and the second one $(\mathrm{A}'_1, \mathrm{A}'_2, \mathrm{A}'_3)$ by the positions of the corresponding three atoms in the target system. In general, the atomic triads are identical for all the ELMOs that belong to a given fragment. The exact definition of the atoms belonging to a triad depends on the type of fragment:

- For a one-atom fragment the atomic triad is defined by the atom on which the corresponding ELMOs are localized and by two additional atoms, which are generally bonded to the first atom.
- The triads for two-atom fragments are constructed using the coordinates of the two atoms included in the fragment and a third atom, which should be representative for an eventually occurring unsymmetrical bonding situation.
- The triad for a three-atom fragment is automatically constructed using the coordinates for all three atoms in the fragment.

The choice of the atomic triads leads to the definition of two reference frames, where the first one $(\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{d})$ is associated with the model molecule and the second one $(\boldsymbol{a}', \boldsymbol{c}', \boldsymbol{d}')$ corresponds to the target system. To define the vectors in the reference frame $(\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{d})$, first the position vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are determined. They respectively give the positions of atoms A_2 and A_3 relative to atom A_1. Vectors $\boldsymbol{c}$ and $\boldsymbol{d}$ are then defined by the following vector products:

$$\begin{cases} \boldsymbol{c} = \boldsymbol{a} \times \boldsymbol{b} \\ \boldsymbol{d} = \boldsymbol{c} \times \boldsymbol{a} \end{cases} \tag{1.29}$$

The reference frame for the target system $(\boldsymbol{a}', \boldsymbol{c}', \boldsymbol{d}')$ is defined in a completely analogous way.

The actual rotation is composed of two transformation steps. In the first step, an ELMO is rotated from its original reference frame $(\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{d})$ to an orthonormal frame $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$, followed by the rotation from $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$ to the target reference frame $(\boldsymbol{a}', \boldsymbol{c}', \boldsymbol{d}')$ in the second step (see Figure 1.6).

Since each generic vector $\boldsymbol{k}$ can be expressed in the orthonormal reference frame $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$ using:

$$\boldsymbol{k} = k_x \hat{\boldsymbol{x}} + k_y \hat{\boldsymbol{y}} + k_z \hat{\boldsymbol{z}} \tag{1.30}$$

the two rotation matrices $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ can be defined for each of the two steps, respectively. Matrix $\boldsymbol{P}_1$ is a directional cosine matrix, which is associated with the rotation from $(\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{d})$

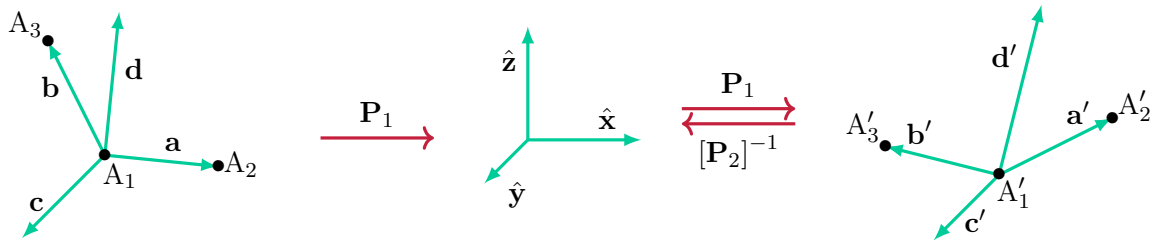


Figure 1.6: Definition of the atomic triads, the corresponding reference frames and the rotation matrices for the rotation of the ELMOs from the geometry of the model molecule to the geometry of the target system.

to $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$:

$$\boldsymbol{P}_1 = \begin{pmatrix} \frac{d_x}{|\boldsymbol{d}|} & \frac{d_y}{|\boldsymbol{d}|} & \frac{d_z}{|\boldsymbol{d}|} \\ \frac{c_x}{|\boldsymbol{c}|} & \frac{c_y}{|\boldsymbol{c}|} & \frac{c_z}{|\boldsymbol{c}|} \\ \frac{a_x}{|\boldsymbol{a}|} & \frac{a_y}{|\boldsymbol{a}|} & \frac{a_z}{|\boldsymbol{a}|} \end{pmatrix} \tag{1.31}$$

while matrix $\boldsymbol{P}_2$ is the inverse[4] of the matrix correponding to the rotation from $(\boldsymbol{a}', \boldsymbol{c}', \boldsymbol{d}')$ to $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$:

$$\boldsymbol{P}_2 = \begin{pmatrix} \frac{d'_x}{|\boldsymbol{d}'|} & \frac{c'_x}{|\boldsymbol{c}'|} & \frac{a'_x}{|\boldsymbol{a}'|} \\ \frac{d'_y}{|\boldsymbol{d}'|} & \frac{c'_y}{|\boldsymbol{c}'|} & \frac{a'_y}{|\boldsymbol{a}'|} \\ \frac{d_z}{|\boldsymbol{d}|} & \frac{c'_z}{|\boldsymbol{c}'|} & \frac{a'_z}{|\boldsymbol{a}'|} \end{pmatrix} \tag{1.32}$$

Finally, the rotation matrix $\boldsymbol{P}$ can be computed from the product of the two matrices above:

$$\boldsymbol{P} = \boldsymbol{P}_2\boldsymbol{P}_1 = \begin{pmatrix} P_{xx} & P_{xy} & P_{xz} \\ P_{yx} & P_{yy} & P_{yz} \\ P_{zx} & P_{zy} & P_{zz} \end{pmatrix} \tag{1.33}$$

Now that transformation matrix $\boldsymbol{P}$ has been constructed, it is possible to define individual rotation matrices for the different types of atomic orbitals (s, p, etc). Since s orbitals are completely spherical, they are invariant to rotations. Thus the transformation matrix for the s functions and the corresponding ELMO coefficients is given by

$$\boldsymbol{S} = S_{11} = 1 \tag{1.34}$$

The p orbitals and their coefficients can be rotated directly using matrix $\boldsymbol{P}$ in the form given by Equation (1.33). Any basis functions with higher angular momentum and the corresponding coefficients cannot be rotated using matrix $\boldsymbol{P}$ directly, but the required matrices may be expressed in terms of matrix $\boldsymbol{P}$.[119]

The matrices $\boldsymbol{S}, \boldsymbol{P}, \boldsymbol{D}$, etc. for the rotation of the basis functions s, p, d, etc. (and the relative ELMO coefficients), can be assembled to a final rotation matrix $\boldsymbol{R}$, which allows the rotation of all the basis functions and coefficients associated with one ELMO. The rotation

[4] Since rotation matrices are orthogonal, the inverse of the matrix is equivalent to its transpose.

matrix $\boldsymbol{R}$ has a block-diagonal structure, because only those coefficients that correspond to basis functions on the same atom and of the same type are combined during the rotation. For an example of the rotation matrix $\boldsymbol{R}$ see Figure 1.7. For the transfer of all the ELMOs from a model to the target system, the different rotation matrices $\boldsymbol{R}$ are combined to a rank-three tensor.

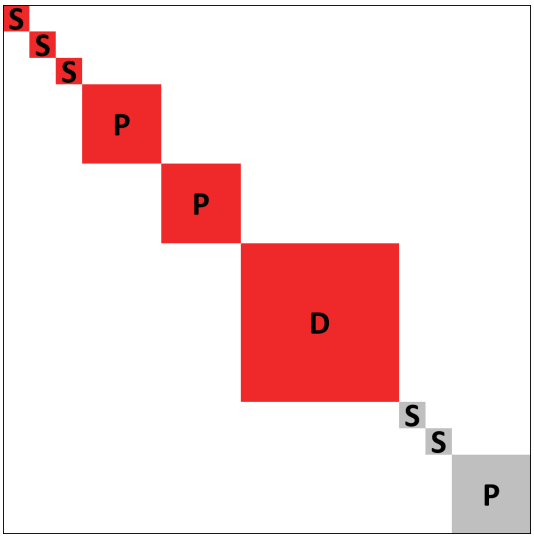


Figure 1.7: Example of a rotation matrix $\boldsymbol{R}$ for an ELMO associated with the O−H bond in the water molecule. The ELMO was computed using the basis set cc-pVDZ. The red and gray squares correspond to the rotation matrices for the basis functions and the corresponding coefficients of the oxygen and hydrogen atom, respectively. For the cc-pVDZ basis set the matrices $\boldsymbol{S}, \boldsymbol{P}$ and $\boldsymbol{D}$ have the dimensions 1×1, 3×3 and 6×6, respectively. Thus, the complete rotation matrix $\boldsymbol{R}$ has a dimension of 20×20.

The transferability of ELMOs has been investigated in different studies for small molecules,[103,104,117,120,131,132] for the frontier bonds in QM/MM calculations[133] and also for large polypeptides and proteins.[119,134] In particular, electron densities obtained after the transfer of ELMOs have been compared to densities resulting from corresponding HF or DFT calculations and from transfers of pseudoatoms that are normally used in crystal structure refinements of crystal structures (see Section 2.3.2). It was shown that the transferability of ELMOs is as reliable as the one of pseudoatoms,[134] and that the electron densities obtained through the transfer of ELMOs are completely comparable to the HF and DFT ones.[119,134]

1.3.4 Libraries of ELMOs

To take advantage of the intrinsic transferability of the extremely localized molecular orbitals, libraries of ELMOs have been recently assembled.[105] Previously, another database of ELMOs (*DENPOL*)[135] had already been constructed. However, it was based only on the minimal basis set STO-4G and the strategy to transfer the ELMOs was slightly different and less reliable than the procedure described above.[105,119] Therefore, the new database was constructed with the goal of "refining crystallographic structures and computing approximate properties of very large molecules",[105] in particular of polypeptides and proteins.

All the possible fragments of the twenty natural amino acids, selenomethionin, and water are currently stored in the database. For the amino acids, all their possible protonation states

and forms (N-terminal, C-terminal or non-terminal) are taken into account. According to their Lewis structures, the amino acids are mostly divided into fragments with one or two atoms. Three-atom fragments are used to describe delocalized units in the structures, for example peptide bonds, carboxylate groups and aromatic rings. In principle, it would be possible to use larger fragments, which could be favorable especially for the description of aromatic rings. However, the libraries were initially constructed for the refinement of crystallographic structures, where aromatic rings are not necessarily constrained to be completely planar, and a misfit between the experimental structures and the theoretically optimized, delocalized ELMOs might occur. In addition, the rotation of ELMOs, which are delocalized on more than three atoms, requires special consideration because it is not possible to define a rotational triad for more than three atoms. Therefore, it would be necessary to store the delocalized ELMOs for all the possible dihedral angles.[105] In fact, this is the adopted procedure to properly transfer the ELMO localized on the S−S bond between two cysteine residues with respect to the lone pairs centered on the sulfur atoms.

To compute the ELMOs for each of the described fragments, model molecules were designed exploiting the nearest functional group approximation, according to which each model molecule consists not only of the particular fragment under exam, but also of the nearest functional groups capped with hydrogen atoms. In a preliminary study of the ELMO transferability,[119] the nearest functional group approximation was found to be the most reliable one for all test cases except for the peptide bond, for which the nearest bond approximation is exploited instead. Hence, the model molecule for the peptide bond consists of the peptide bond and its nearest neighboring bonds, which are saturated with hydrogen atoms. Following the procedure described in Section 1.3.2, the ELMOs in the libraries are computed on optimized structures of the model molecules. Currently, the occupied and virtual ELMOs for all the fragments of the twenty-one amino acids and water are stored in the ELMO libraries together with their associated atomic triads.[105] The ELMOs in the databanks are available for five standard basis sets, namely 6-31G,[136–140] 6-311G,[141–143] 6-31G(d,p),[136–140,144] 6-311G(d,p),[141–143] and cc-pVDZ.[145–147]

The ELMO libraries are associated with a stand-alone software that is also called *ELMOdb* program.[105] This program requires as minimum input a PDB file in AMBER[148] format. In addition, also an xyz file may given, if a larger precision is required for the coordinates of the atoms. This is for example the case in crystallographic refinements.[149] However, a PDB file is always necessary, because it allows the *ELMOdb* program to reliably assign the ELMOs to the different amino acids and the corresponding atoms. It is important to note that the PDB file must include hydrogen atoms,[105] but it may contain other residues than the twenty-one amino acids and water. In fact, the *ELMOdb* program offers the possibility to add tailor made residues. For these residues, the externally computed ELMOs can be read by the software, also if they were calculated with a different basis set than the five ones currently available.

For all the fragments associated with all the residues in the PDB file, the program constructs a rotation matrix (as described in Section 1.3.3) and instantaneously transfers all the orbitals. Afterwards, the transferred ELMOs are renormalized to take into account differences in the bond lengths and angles between the model and target structures. Because the transferred ELMOs are non-orthogonal to each other after the transfer, they may be orthog-

onalized by the program. In this way, the wavefunction of the target system is reconstructed and is provided to the user either as a Gaussian[150] formatted checkpoint file (fchk format) or a wavefunction file (wfx format), which offers the possibility to use the ELMO wavefunction for example in the refinement of crystal structures and for the analysis of non-covalent interactions.

1.4 Embedding methods

An alternative to the previously discussed fragmentation approaches is represented by the embedding techniques. In contrast to the former, in the latter techniques the target system is divided into two main subsystems: a region of interest A and its environment B. There may be more than one region of interest and also the environment may be further divided into different layers.[151] After the division of the system, the goal is to calculate properties of region A using quantum mechanics (QM) methods. Since these properties are influenced by the environment, region B is also included in the calculations.

In case of large systems, for example proteins, this can mean that several thousand atoms need to be treated simultaneously, while, at the same time, the chemical process of interest must be accurately described.[152] Therefore, in multiscale embedding techniques, different levels of theory are used together. For example, in the well-known quantum mechanics / molecular mechanics (QM/MM) approach[152,153] the chemically active region is described using QM techniques, while its surroundings are treated by means of MM methods. In continuum solvent models one component (typically the solute) is described at QM level, while its environment (typically the solvent) is treated using a responsive continuum medium.[154–158] Alternatively, fully QM embedding approaches apply only QM methods with different levels of accuracy: a high level is used for the region of interest, and a lower one for the surroundings.[159]

In Section 1.4.1, the main ideas of the QM/MM strategy will be summarized, followed by an introduction to fully QM embedding techniques in Section 1.4.2.

1.4.1 The QM/MM embedding technique

The QM/MM approach was first proposed by Warshel and Levitt in 1976, who applied the technique to study an enzymatic reaction of lysozyme.[153] In 2013, the Nobel prize was awarded to Warshel, Levitt and Karplus "for the development of multiscale models for complex chemical systems."[160–163] Nowadays, the QM/MM technique has become a widely applied and popular method for the study of large systems, such as biomolecules, inorganic and organometallic systems as well as solid state systems, but also to study explicit solvation processes.[152]

As mentioned above, traditional QM and MM methods are combined in the QM/MM technique. MM force fields[164] are fast and effective in simulating large biosystems, but they cannot be used to describe chemical reactions. In contrast, QM methods are developed to study chemical systems and their reactivity, but their application is often limited to small molecules (see Section 1.1.4). Therefore, in the QM/MM approach, the chemically active region is described using QM techniques, while the surroundings are treated by means of MM methods. In principle, almost any combination of MM and QM techniques is possible. In

practice, mostly DFT is used for the description of the QM part.[152,165]

Within the QM/MM approach two ways of computing the total energy of the system exist, namely, an additive or subtractive scheme.[152,166] In the former scheme, the total energy of the QM/MM system is composed of the energy E^A_{QM} obtained from a QM calculation on the region of interest A and the energy E^B_{MM} from an MM calculation on the environment B plus an energy $E^{AB}_{QM/MM}$ accounting for the coupling between the two subsystems:

$$E_{additive} = E^A_{QM} + E^B_{MM} + E^{AB}_{QM/MM} \tag{1.35}$$

In the right-hand side of Equation (1.35) the subscripts indicate the type of the calculation, while the superscripts denote the different subsystems, with A being the QM region, B the MM subsystem and AB the combination of both subsystems. In contrast, in the subtractive scheme, three individual energies are calculated by performing one MM calculation on the complete system (E^{AB}_{MM}), a second MM calculation on the QM region (E^A_{MM}) and a QM calculation on the QM region (E^A_{QM}). The total energy according to the subtractive scheme is then given by:

$$E_{subtractive} = E^{AB}_{MM} - E^A_{MM} + E^A_{QM} \tag{1.36}$$

Depending on the target system, the regions A and B may be defined in such a way that cutting through covalent bonds is avoided (for example in solvation studies).[152] However, in many systems the frontier between the QM and MM parts cuts through covalent bonds. In analogy with the previously discussed fragmentation techniques, also in the QM/MM approach it is necessary to saturate the dangling bonds of the atoms at the frontier to achieve a well-balanced description of the interactions between the QM and MM parts.[152] To accomplish this task, three different boundary schemes are available.[152] The first option is to saturate the free valencies at the boundary of the QM system using link atoms, which are in most of the cases hydrogen atoms that are not part of the original system. As a second option, the atoms at the MM frontier may be replaced by special boundary atoms. Finally, localized orbitals may be transferred to the frontier bonds and some of these orbitals are kept frozen in the SCF calculations. One representative example for the last group is the local self-consistent field (LSCF) method,[167–170] where the SCF calculation for the QM region is performed in the presence of frozen localized MOs. The latter are either standard localized MOs, from which the tails have been deleted,[170] but also ELMOs can be used.[133]

The interaction between the QM and MM parts may be described with different embedding schemes.[152,165] In the mechanical embedding, the electrostatic interaction is computed by using the identical charge model in the QM and MM parts. Therefore, the description of the interactions is entirely at the MM level. In contrast, in the electrostatic embedding, the QM calculation is performed in the presence of the MM charges. This scheme has the advantage that the interaction is described more accurately because the QM part is polarized by the MM part. To also include polarization in the opposite direction (from the QM to the MM region), mutual polarization may be included with the polarized embedding methods that take advantage of polarizable force fields.

A popular example of a QM/MM strategy is the ONIOM technique[171,172] (ONIOM stands for our Own N-layer Integrated molecular Orbital molecular Mechanics), which is based on

the subtractive scheme. As the name suggests, ONIOM is not restricted to two layers. For example, a three-layer variant has been developed by Svensson and coworkers,[171] where the complete system is divided into a model, an intermediate and a real layer. The model layer is described at high level using QM, in particular *ab initio* methods or DFT, while for the intermediate layer a lower level *ab initio*, DFT or semiemperical method is used. The real layer is then described using a semiempirical method or MM.[172] If all three levels are described using QM methods, the ONIOM approach also falls into the group of the fully QM multilevel techniques, which will be described in the next subsection.

In the context of this thesis, it is interesting to note that the experimental data in protein crystal structure refinements are typically supplemented by empirical restraints (for details about restraints, see Section 2.1.9), which can take forms of MM force fields.[173] These restraints are usually accurate for amino acids and nucleic acids but not for metal sites and other parts of the protein.[174,175] To improve their description, the quantum refinement technique[174–177] has been developed. In the original version of this strategy, the protein region of interest is described at QM level and its surroundings at MM level. However, other versions exist as well.[174] For example, one particular quantum refinement approach makes use of the "divide & conquer" strategy (see Section 1.2.1) to describe the whole system by means of QM.[178] Quantum refinement techniques have been applied successfully to obtain more accurate protein crystal structures.[174]

1.4.2 Fully QM strategies

While in the previously discussed QM/MM approach the QM part is embedded in a subsystem described by means of classical techniques, in the quantum embedding methods the complete system is described using quantum chemical strategies: a high level one for the subsystem of interest and a low level one for its environment. Several quantum embedding approaches have been developed in the past years. Three great families are the density functional embedding techniques, the density matrix embedding methods and the Green's function embedding approaches.[151] A detailed description of all three families is beyond the scope of this thesis. Therefore, only the key ideas of three density functional embedding techniques will be summarized here, in a similar fashion as it was previously done for the fragmentation approaches. In particular, the three techniques that will be described in more detail are the frozen density embedding theory (FDET), the projection-based embedding (PbE) strategy and the mutilevel density functional theory (MLDFT) approach. The reason for choosing these methods is that they are the closest to the QM/ELMO approach (see Section 1.5), which is the fully QM embedding strategy used in the research work described in this thesis.

Before coming to the details of the different strategies, let us first consider some general aspects that were in this form also pointed out by Manby in reference [179]. If in KS-DFT a system is divided into regions A and B, then the electron density can be partitioned accordingly:

$$\rho(\boldsymbol{r}) = \rho^A(\boldsymbol{r}) + \rho^B(\boldsymbol{r}) \tag{1.37}$$

For this partitioning, the total Kohn-Sham energy takes the following form:

$$E[\rho] = E[\rho^A] + E[\rho^B] + \delta E[\rho^A, \rho^B] \tag{1.38}$$

where the first terms are related only to the corresponding region A or B and the last term consists of all the nonadditive contributions to the energy. In particular, the latter contains contributions for the Hartree energy (which is known), for the exchange-correlation (which needs to be approximated as described in Section 1.1.3) and for the kinetic energy. Depending on the chosen partitioning, the computation of the latter can be a challenge. If the densities for regions A or B are constructed from mutually orthogonal KS orbitals, then the kinetic energy of the complete system is given by:

$$T_s[\rho] = T_s[\rho^A] + T_s[\rho^B] \tag{1.39}$$

However, as we will see below, in FDET the orbitals are obtained separately for the two subsystems, leading to the appearance of the non-additive kinetic energy functional that is defined using:

$$T_s^{nadd}[\rho^A, \rho^B] = T_s[\rho] - T_s[\rho^A] - T_s[\rho^B] \tag{1.40}$$

which needs to be approximated.

In the following, the FDET is an example for a technique where the non-additive kinetic energy functional is present, while the PbE and MLDFT methods can avoid this term.

The frozen density embedding theory (FDET)[180–182] is a technique developed by Wesołowski *et al.*, where the target system is divided into an active region A and a frozen environment B. The total density of the system is separated into two contributions ρ^A and ρ^B according to Equation (1.37). The underlying idea of FDET is to embed the active system A in an effective environmental potential, which can be derived from the density of the environment $\rho^B(\boldsymbol{r})$.[183] In the beginning of the calculation, an approximation is made to the density $\rho^B(\boldsymbol{r})$, for example by further subdividing system B into smaller fragments and summing up the contributions for the individual fragment densities.[182] Then, the density $\rho^A(\boldsymbol{r})$ of the active system can be calculated in the presence of the frozen density of system B.[183] In practice, this is done solving the following set of modified Kohn-Sham equations:

$$\left[-\frac{1}{2}\boldsymbol{\nabla}^2 + v_{KS}^A[\rho_A](\boldsymbol{r}) + v_{emb}[\rho_A, \rho_B](\boldsymbol{r})\right] \varphi_{iA}(\boldsymbol{r}) = \epsilon_{iA}\varphi_{iA}(\boldsymbol{r}) \tag{1.41}$$

where v_{KS}^A is the Kohn-Sham potential for the isolated region A and v_{emb}^K is the effective embedding potential representing the influence of region B on A. The latter depends on the functional derivative of the non-additive kinetic energy functional, which needs to be approximated since the KS orbitals are determined separately for the two regions but not for the complete system.[180,184] An accurate approximation of the non-additive kinetic energy functional remains challenging although desirable, especially for the description of systems with covalent bonds between regions A and B.[182]

A FDET variant that allows the fully self-consistent polarization of the subsystem densities is realized using a "freeze and thaw" procedure, where the active and the frozen subsystems change their roles iteratively, until the densities of all the subsystems converge.[183,185] Furthermore, FDET has been generalized to compute excitation energies and absorption spectra using time-dependent DFT.[186,187]

Applications of FDET[182] include for example the study of chromophors in biomolecules[188] and solvent effects on different molecular properties like dipole and quadrupole moments[184] as well as electronic excitation energies[184,189,190] and electronic absorption spectra[190].

The projection-based embedding (PbE) method[179,191] provides exact DFT-in-DFT embedding and approximate wavefunction-in-DFT (WF-in-DFT) embeddings.[191]

The general procedure of the PbE technique starts with a low level Kohn-Sham DFT calculation on the complete system. Afterwards, the occupied Kohn-Sham orbitals are localized and grouped into two subsystems A and B, where subsystem A denotes the region of interest, while its surroundings are included in subsystem B. In PbE, a level shift projection operator is applied, which enforces the orbitals in subsystem A to be orthogonal to those in subsystem B. Therefore, as was mentioned above, the problem of evaluating the nonadditive kinetic energy functional is avoided, since the kinetic energy is given by Equation (1.39). In addition, the projection operator shifts the orbitals of subsystem B to higher energies (compared to the orbital energies of the subsystem A). In this way, it is ensured that the orbitals in subsystem B cannot hybridize with those in subsystem A.[191] For the embedded subsystem A, iterative SCF calculations are performed, exploiting either higher level DFT functionals (DFT-in-DFT embedding) or correlated wavefunction methods (WF-in-DFT embedding). The latter is implemented in such a way, that any wavefunction method may be used without additional programming.[191]

The localized orbitals in the standard PbE method are obtained using the Pipek-Mezey technique[112] (see Section 1.3.1) or similar strategies.[179,191] Therefore, the orbitals of one subsystem are also expanded on basis functions that are centered on atoms of the other subsystem. In contrast, in the absolutely localized variant of PbE,[192] the orbitals of one subunit use only basis functions that are centered on atoms belonging to that subunit. The absolutely localized PbE method reduces the computational cost of the wavefunction calculation because the number of orbitals in the wavefunction region is reduced.[192]

The WF-in-DFT PbE method has been used to investigate for example transition-metal catalysis, enzyme catalysis and battery electrolyte decomposition.[191] Although the technique was mostly applied to molecular systems, it was also employed to study periodic systems with[193] and without[194] the absolutely localized variant. Furthermore, the PbE approach has been also coupled with the QM/MM method for investigating enzyme reactions.[195,196]

The mutilevel density functional theory (MLDFT) approach[197] is also based on the decomposition of the target system into an active and inactive region. More precisely, the initial density matrix for the complete system is constructed from a superposition of fragment densities[197] and subsequently partitioned into two subsystem density matrices:

$$\boldsymbol{P} = \boldsymbol{P}^A + \boldsymbol{P}^B \tag{1.42}$$

where A denotes the active fragment and B the inactive one. In practice, the partitioning is performed using Cholesky decomposition,[198] which generates a set of active occupied orbitals. From this set of orbitals the active density matrix $\boldsymbol{P}^A$ is calculated. The embedding density matrix is the remainder of the initial density matrix after the decomposition:

$\boldsymbol{P}^B = \boldsymbol{P} - \boldsymbol{P}^A$. The density matrix $\boldsymbol{P}^B$ is computed only once at the beginning, because it remains frozen during the following SCF calculations, providing an external field for the active region. The Cholesky decomposition procedure ensures that all the orbitals in the active and inactive regions are mutually orthogonal to each other, which avoids the necessity of computing the non-additive kinetic energy functional in the MLDFT approach. After the Cholesky decomposition, the Kohn-Sham equations are solved for the active MOs only, while the density matrix of the inactive region remains frozen, thus maintaining the orthogonality of the two parts during the SCF procedure.

The MLDFT technique is similar to the above described FDET. However, in FDET the electron density is partitioned, while in MLDFT the density matrix is decomposed, thus, the problem of the nonadditive kinetic potential terms is avoided in MLDFT. Furthermore, if the decomposition would be performed on the fully converged DFT density matrix, the MLDFT energy would correspond to the exact DFT energy of the full system. The same result may also be obtained with the previously described PbE technique. Nevertheless, compared to PbE, MLDFT does not require a full DFT calculation on the complete system because the initial density matrix is constructed simply from the sum of individual fragment densities.

Other than the recently developed MLDFT, different variants of the multilevel approach exist, such as the mutilevel Hartree-Fock (MLHF)[199,200] and multilevel coupled cluster (MLCC)[201–203] techniques. Recently, both variants have been combined with the QM/MM approach to study excitation energies of molecular systems in aqueous solution.[204]

1.5 Embedding techniques based on extremely localized molecular orbitals

This section is dedicated to the description of the embedding method that is fundamental for the research work presented in this thesis, namely the quantum mechanics / extremely localized molecular orbital (QM/ELMO) technique.[205–207] Following an analogous strategy as the previously described embedding approaches, the QM/ELMO technique is based on a partitioning of the target system into a region of interest and its environment. In the QM/ELMO method, the former is described with traditional QM methods, while the latter is treated by means of frozen ELMOs (see Section 1.3).

The QM/ELMO method has been developed by modifying the previously mentioned LSCF approach[167–170] (see Section 1.4.1). In the original version of the QM/ELMO technique only QM calculations at HF level were possible.[205] Later on, it has been further extended to DFT and post-HF strategies[206] and to methods for the investigation of excited states.[207–209] Very recently, the QM/ELMO approach has been also combined with molecular mechanics giving rise to the three-layer quantum mechanics / extremely localized molecular orbital / molecular mechanics (QM/ELMO/MM) multiscale technique.[210]

This section is organized as follows. The SCF procedure of QM/ELMO calculations at HF or DFT level will be explained in Section 1.5.1. Henceforth, these two types of QM/ELMO calculations are abbreviated with the acronyms HF/ELMO and DFT/ELMO, respectively. Finally, the most important aspects of the QM/ELMO/MM technique will be described in Section 1.5.2.

1.5.1 The QM/ELMO technique

The first step of any QM/ELMO calculation is the partitioning of the target system into a QM and an ELMO region (see Figure 1.8). In case the frontier between the two regions cuts through covalent bonds, the two regions share only the frontier atoms and only the basis functions centered on these atoms are common to the basis sets of both regions. In the second step, the ELMOs are transferred to the ELMO region, following the strategy described in Section 1.3.3.

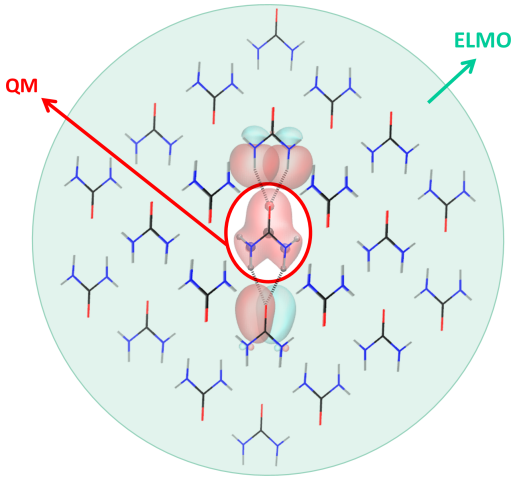


Figure 1.8: Example for the definition of the different regions in a QM/ELMO calculation on a cluster of urea molecules. The central molecule is placed in the QM region, while the remaining molecules in the cluster are part of the ELMOs region. Adapted with permission from reference [207]. Copyright 2021 American Chemical Society.

Before the SCF iterations can start, it is first necessary to orthogonalize the orbitals. The complete orthogonalization procedure consists of three steps:

1. The transferred ELMOs are orthogonalized using the Löwdin orthonormalization, which mostly preserves the localized nature of the ELMOs;[206]
2. the QM basis functions are orthogonalized against the orthonormalized ELMOs that were obtained from the preceding step;
3. the QM basis functions resulting from the previous step are canonically orthogonalized.

All three steps can be summarized with the transformation:

$$\boldsymbol{\chi}' = \boldsymbol{\chi}\boldsymbol{B}, \tag{1.43}$$

where $\boldsymbol{\chi}$ is the starting array of the M nonorthogonal basis functions for the complete system with dimensions $1 \times M$:

$$\boldsymbol{\chi} = [|\chi_1\rangle, |\chi_2\rangle, \ldots, |\chi_M\rangle] \tag{1.44}$$

$\boldsymbol{\chi}'$ is the final $1 \times M_{\mathrm{QM}}$ array for the M_{QM} orthonormal basis functions of the QM region:

$$\boldsymbol{\chi}' = \left[|\chi'_1\rangle, |\chi'_2\rangle, \ldots, \left|\chi'_{M_{\mathrm{QM}}}\right\rangle\right] \tag{1.45}$$

and $\boldsymbol{B}$ is a transformation matrix with dimensions $M \times M_{\mathrm{QM}}$, which plays a central role in the QM/ELMO self-consistent field algorithm, as will be described below.

After the previously described steps have been performed, the actual SCF calculation starts. First, the Fock matrix $\boldsymbol{F}$ is constructed in the original basis set $\boldsymbol{\chi}$. In the QM/ELMO procedure, the elements of the Fock matrix can be separated into four individual contributions:

$$\begin{aligned} F_{\mu\nu} &= h_{\mu\nu} + F_{\mu\nu}^{QM} + F_{\mu\nu}^{ELMO} + v_{\mu\nu}^{XC} \qquad (1.46) \\ &= \langle\chi_\mu|\hat{h}^{core}|\chi_\nu\rangle \\ &+ \sum_{\lambda,\sigma=1}^{M} P_{\lambda\sigma}^{QM}\left[(\chi_\mu\chi_\nu|\chi_\sigma\chi_\lambda) - \frac{1}{2}x\,(\chi_\mu\chi_\lambda|\chi_\sigma\chi_\nu)\right] \\ &+ \sum_{\lambda,\sigma\in ELMO} P_{\lambda\sigma}^{ELMO}\left[(\chi_\mu\chi_\nu|\chi_\sigma\chi_\lambda) - \frac{1}{2}x\,(\chi_\mu\chi_\lambda|\chi_\sigma\chi_\nu)\right] \\ &+ \langle\chi_\mu|\hat{v}^{XC}\left[\boldsymbol{P}^{QM} + \boldsymbol{P}^{ELMO}\right]|\chi_\nu\rangle \qquad (1.47) \end{aligned}$$

where $\hat{h}^{core}$ is the standard core one-electron Hamiltionian operator, $\boldsymbol{P}^{QM}$ and $\boldsymbol{P}^{ELMO}$ are the one-electron reduced density matrices in the original basis set $\boldsymbol{\chi}$ for the QM and ELMO parts, respectively. The two-electron repulsion integrals are denoted with $(\chi_\alpha\chi_\beta|\chi_\gamma\chi_\delta)$ and x is the fraction of exact exchange. Finally, the fourth term in Equation (1.46), $v_{\mu\nu}^{XC}$, is the element (μ,ν) of the exchange-correlation potential matrix, which depends on the global one-electron reduced density matrix $\boldsymbol{P}$ that is given by the sum of the two density matrices $\boldsymbol{P}^{QM}$ and $\boldsymbol{P}^{ELMO}$. In the case of a HF/ELMO calculation, the fraction of exact exchange x is equal to 1 and the exchange-correlation potential ($v_{\mu\nu}^{XC}$) disappears from the equations. In the DFT/ELMO approach, the transferred and orthonormalized ELMOs are treated as frozen Kohn-Sham orbitals.

The complete Fock matrix needs to be constructed only once at the beginning of the SCF procedure. In the actual SCF iterations, only the contribution from the QM subsystem ($F_{\mu\nu}^{QM}$) and, in case of DFT/ELMO, the exchange-correlation term ($v_{\mu\nu}^{XC}$) change, while the contributions from the one-electron Hamiltonian operator and from the ELMO regions remain constant. This is because the SCF iterations are performed only for the QM part, while the ELMO region remains frozen. Therefore, in the next step, the complete Fock matrix $\boldsymbol{F}$ is transformed to the Fock matrix $\boldsymbol{F'}$ for the QM subsystem in the basis set $\boldsymbol{\chi'}$ using again the transformation matrix $\boldsymbol{B}$ and its transpose $\boldsymbol{B}^\top$:

$$\boldsymbol{F'} = \boldsymbol{B}^\top\boldsymbol{F}\boldsymbol{B} \qquad (1.48)$$

While the Fock matrix $\boldsymbol{F}$ has the dimensions $M\times M$, $\boldsymbol{F'}$ is only an $M_{\mathrm{QM}}\times M_{\mathrm{QM}}$ matrix. Since $M >> M_{\mathrm{QM}}$, the transformation in Equation (1.48) significantly reduces the computational cost of the next step, the diagonalization of the Fock matrix $\boldsymbol{F'}$ to solve the Roothaan-Hall equations:

$$\boldsymbol{F'}\boldsymbol{C'} = \boldsymbol{C'}\boldsymbol{E'} \qquad (1.49)$$

After the solutions to Equation (1.49) are obtained, the matrix of the MO coefficients $\boldsymbol{C'}$ is transformed back to the original basis set $\boldsymbol{\chi}$ using again the transformation matrix $\boldsymbol{B}$:

$$\boldsymbol{C} = \boldsymbol{B}\boldsymbol{C'} \qquad (1.50)$$

Finally, the new QM one-electron density matrix in the original complete basis set $\boldsymbol{\chi}$ is computed using the new matrix $\boldsymbol{C}$:

$$P_{\lambda\sigma}^{QM} = 2\sum_{i=1}^{N} C_{\sigma i}^{*} C_{\lambda i} \tag{1.51}$$

for which convergence is checked. If convergence is not reached, the Fock matrix is updated with the new density matrix for the QM subsystem from Equation (1.51), while the ELMO density matrix does not need to be updated since it remains frozen during the SCF iterations. Otherwise, if convergence is achieved, the total energy of the system and other required properties can be calculated. In particular, the energy for a HF/ELMO calculation is given by:

$$\begin{aligned} E_{HF/ELMO} = & \frac{1}{2}\sum_{\mu,\nu=1}^{M} P_{\nu\mu}^{QM}\left(2h_{\mu\nu}^{core} + F_{\mu\nu}^{QM}\right) \\ & + \frac{1}{2}\sum_{\mu,\nu=1}^{M} P_{\nu\mu}^{ELMO}\left(2h_{\mu\nu}^{core} + F_{\mu\nu}^{ELMO}\right) \\ & + \frac{1}{2}\left(\sum_{\mu,\nu=1}^{M} P_{\nu\mu}^{QM} F_{\mu\nu}^{ELMO} + \sum_{\mu,\nu=1}^{M} P_{\nu\mu}^{ELMO} F_{\mu\nu}^{QM}\right) \end{aligned} \tag{1.52}$$

where the matrix $\boldsymbol{h}^{core}$ is associated with the core one-electron Hamiltonian operator and the one-electron density matrices $\boldsymbol{P}^{QM}$ and $\boldsymbol{P}^{ELMO}$ are analogous to those in Equation (1.47). Also the matrices $\boldsymbol{F}^{QM}$ and $\boldsymbol{F}^{ELMO}$ are the same as in Equation (1.46) since the fraction of exact exchange equals 1 at HF level. Equation (1.52) can be decomposed into three contributions, since the first term is a pure QM contribution, the second one a pure ELMO contribution, and the third term is a mixed QM/ELMO contribution.

The total energy for a DFT/ELMO calculation is given by:

$$\begin{aligned} E_{DFT/ELMO} = & \frac{1}{2}\sum_{\mu,\nu=1}^{M} P_{\nu\mu}^{QM}\left(2h_{\mu\nu}^{core} + F_{\mu\nu}^{QM}\right) \\ & + \frac{1}{2}\sum_{\mu,\nu=1}^{M} P_{\nu\mu}^{ELMO}\left(2h_{\mu\nu}^{core} + F_{\mu\nu}^{ELMO}\right) \\ & + \frac{1}{2}\left(\sum_{\mu,\nu=1}^{M} P_{\nu\mu}^{QM} F_{\mu\nu}^{ELMO} + \sum_{\mu,\nu=1}^{M} P_{\nu\mu}^{ELMO} F_{\mu\nu}^{QM}\right) \\ & + E^{XC}\left[\boldsymbol{P}^{QM} + \boldsymbol{P}^{ELMO}\right] \end{aligned} \tag{1.53}$$

Also in Equation (1.53), the matrices $\boldsymbol{h}^{core}$, $\boldsymbol{P}^{QM}$ and $\boldsymbol{P}^{ELMO}$ correspond to the ones in Equation (1.47). The Fock matrices $\boldsymbol{F}^{QM}$ and $\boldsymbol{F}^{ELMO}$ are characterized by a fraction of exact exchange x, which depends on the chosen functional. Furthermore, the last term in Equation (1.53), $E^{XC}\left[\boldsymbol{P}^{QM} + \boldsymbol{P}^{ELMO}\right]$, is the exchange-correlation energy, which depends on the functional and on the global density matrix $\boldsymbol{P} = \boldsymbol{P}^{QM} + \boldsymbol{P}^{ELMO}$. Since the exchange-correlation functional is non linear, it is not possible to separate the QM and ELMO contri-

butions. Therefore, unlike Equation (1.52) for the HF/ELMO energy, Equation (1.53) cannot be decomposed into purely QM and ELMO parts.

The recently developed QM/ELMO approach is implemented in a modified version of *Gaussian09*.[150] The technique has been applied to study chemical reactions, bond dissociations and intermolecular interactions of small molecules,[206] but also to compute a protein-ligand interaction energy[205] and for the description of excited states.[208] Furthermore, a central part of the research work that will be presented in this thesis is based on applications of the QM/ELMO method, in particular to crystal structure refinements and to the study of interactions in large systems (i.e. proteins).

1.5.2 The QM/ELMO/MM technique

In the QM/ELMO/MM technique, the previously described QM/ELMO strategy is combined with the QM/MM approach (compare Section 1.4.1).

In analogy with the QM/ELMO method, a QM/ELMO/MM calculation starts with the partitioning of the system, this time into three regions: a QM, an ELMO and an MM one. The frontiers between the QM and the ELMO layers are treated as in the QM/ELMO strategy. If the frontier between the ELMO and the MM subsystems cuts through covalent bonds, the link atom scheme is applied.

The QM/ELMO/MM approach is based on the additive scheme to compute the total energy:

$$E = E^{A}_{QM/ELMO} + E^{B}_{MM} + E^{AB}_{QM/ELMO/MM} \tag{1.54}$$

where the subscripts indicate again the type of the calculation, while the superscripts denote the different subsystems, with A being the combined QM/ELMO region, B the MM subsystem and AB the combination of all the subsystems. In particular, the first energy $E^{A}_{QM/ELMO}$ is a purely quantum mechanical energy for the combined QM and ELMO regions. The procedure to compute this term will be further discussed below. The second term E^{B}_{MM} in Equation (1.54) is a purely classical energy, which is composed of bonded terms, van der Waals terms and electrostatic interaction terms, obtained from standard force fields. Finally, the third energy $E^{AB}_{QM/ELMO/MM}$ is a hybrid term, which includes different contributions from the interactions of the MM subunit with the QM and ELMO regions. In particular, it comprises electrostatic, van der Waals and also bonding terms (the latter only if covalent bonds between the MM and ELMO regions exist).

The procedure to compute the energy $E^{A}_{QM/ELMO}$ is practically identical to the one for a QM/ELMO calculation (compare Section 1.5.1). In fact, the only difference is that an additional fifth term $F^{MM}_{\mu\nu}$ is added to the the Fock operator in comparison to Equation (1.47), which is a one-electron term that accounts for the electrostatic interactions between the QM/ELMO subsystem and the point charges of the MM region. Thus, the Fock matrix

for a QM/ELMO/MM calculation becomes:

$$
\begin{aligned}
F_{\mu\nu} &= h_{\mu\nu} + F_{\mu\nu}^{QM} + F_{\mu\nu}^{ELMO} + v_{\mu\nu}^{XC} + F_{\mu\nu}^{MM} && (1.55)\\
&= \langle\chi_\mu|\hat{h}^{core}|\chi_\nu\rangle \\
&+ \sum_{\lambda,\sigma=1}^{M} P_{\lambda\sigma}^{QM}\left[(\chi_\mu\chi_\nu|\chi_\sigma\chi_\lambda) - \frac{1}{2}x\,(\chi_\mu\chi_\lambda|\chi_\sigma\chi_\nu)\right] \\
&+ \sum_{\lambda,\sigma\in ELMO} P_{\lambda\sigma}^{ELMO}\left[(\chi_\mu\chi_\nu|\chi_\sigma\chi_\lambda) - \frac{1}{2}x\,(\chi_\mu\chi_\lambda|\chi_\sigma\chi_\nu)\right] \\
&+ \langle\chi_\mu|\hat{v}^{XC}\left[\boldsymbol{P}^{QM} + \boldsymbol{P}^{ELMO}\right]|\chi_\nu\rangle \\
&+ \sum_{K\in MM} \langle\chi_\mu|\frac{q_K}{R_{iK}}|\chi_\nu\rangle && (1.56)
\end{aligned}
$$

The QM/ELMO/MM method has been implemented by coupling the in-house modified version of the *Gaussian09*[150] quantum chemistry software with the molecular dynamics package *AMBER 2016*[148]. The former handles the quantum contributions, while the latter deals with the classical ones. In a first study, the QM/ELMO/MM technique has been used to study an enzyme reaction.[210] Furthermore, in the research work presented in this thesis, it was applied in the refinement of crystal structures.

1.6 Summary of the introduction and outlook to the next parts of the thesis

Coming back to the beginning of this chapter, as Dirac pointed out approximately 100 years ago, the fundamental laws for a mathematical description of chemistry are available. Furthermore, as outlined in Section 1.1, nowadays we have the choice between several quantum chemical methods that allow us to approximately solve the equations that he was referring to. However, having in mind the computational cost of these methods, which was described in Section 1.1.4, we can now also understand that Dirac was very optimistic when he declared that "*the main features of complex atomic systems*" could be explained "*without too much computation.*"[1]

Therefore, different techniques were specifically developed for the application to large systems, in particular fragmentation and embedding techniques. The former have been described in Section 1.2. In summary, the main idea of this type of methods is to divide a large system into more manageable fragments, to compute a certain property for the individual fragments, and to recombine the results for all the fragments with the goal of approximating the property of the complete system. Furthermore, several embedding techniques have been described in Section 1.4. Instead of dividing the complete system into several subunits, which are treated at the same level of accuracy, the embedding techniques account for many cases in which one region of a system is particularly of interest, while its environment is only relevant because it influences the properties of the interesting region. Therefore, the target system is divided into two main subsystems in the embedding techniques: a system of interest and its environment, which are described using different levels of accuracy.

For both types of strategies, some selected techniques have been briefly reviewed. In

contrast, in Sections 1.3 and 1.5, the ELMO libraries and QM/ELMO embedding method were described in detail because they are fundamental to the research work that will be described in the next two parts of this thesis. In particular, both techniques were applied for the refinement of crystal structures (see Part I) and for the analysis of non-covalent interactions (see Part II).

Part I

Crystal structure refinement based on quantum mechanical methods

2 Introduction to the refinement of crystal structures and beyond

Single-crystal X-ray diffraction experiments are an indispensable tool for determining the three dimensional structure of small and large molecules routinely and reliably.[211–213] Historically, several structure models have been significantly influenced by the results of X-ray diffraction experiments, especially those of biomolecules such as proteins[212,214] and DNA[215–217]. For example, when the first protein structure of myoglobin[218] was determined by means of X-ray crystallography, its structure appeared to be very complex and irregular. That came as a quite unpleasant surprise to several scientists in the 1960s, who had imagined the structure to be much simpler and more regular.[214] To this day, X-ray diffraction remains the most common technique for determining the three dimensional structures of small and large molecular systems.[211] For example, this can be evinced in Figure 2.1 that shows the number of entries in the PDB,[219] where the majority of structures has been determined by means of X-ray crystallography.

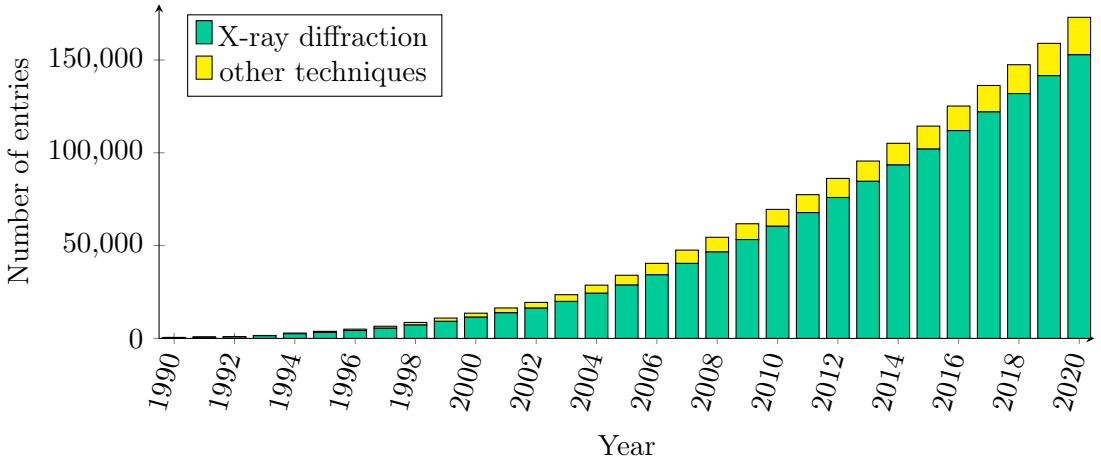


Figure 2.1: Total number of entries and corresponding number of X-ray structures in the protein data bank from 1990 to 2020. Data from `www.rcsb.org`, accessed on May 19, 2021.

Nevertheless, it is important to note that X-ray diffraction experiments yield information that go beyond the positions of non-hydrogen atoms.[175] X-rays are scattered by electrons, and the corresponding diffraction experiment can give valuable insights into the electron density distribution in the crystal. In fact, the atomic positions in the crystal are only indirectly (but very reliably) determined by means of X-ray diffraction, since the maxima in the electron densities correspond in most cases to the nuclear positions of non-hydrogen atoms. However, if high-quality and high-resolution X-ray data are collected, information about electron density regions that are associated with chemical bonding is obtained and needs to be modeled. To accomplish this task, the standard model of X-ray crystallography,

the IAM,[220] is insufficient because it is based on averaged spherical densities that cannot account for the aspherical deformation of the electron density in molecules. Moreover, with the IAM only the positions of non-hydrogen atoms can be determined reliably, while hydrogen atoms are systematically shifted into the bond, leading to too short bond lengths involving hydrogen atoms. However, accurate positions of hydrogen atoms are important in many cases,[149,221] for example for the correct description of intermolecular interactions such as hydrogen bonds or agositc interactions.[222–224] Therefore, it is highly desirable to establish more advanced models that describe the features of chemical bonding and that yield more accurate structures. This is one of the main goals of quantum crystallography.[175,225]

In this context, it is important to note that there has always been a close connection between crystallographic models and theoretical chemistry.[225,226] In fact, as was already mentioned in Section 1.1, the electron density is a physical observable that can be calculated theoretically and derived experimentally. Therefore, scattering experiments[1] and theoretical calculations are complementary approaches that can both give access to the electronic structure of molecules.[175] The influence of crystallography on theoretical chemistry used to be rather indirect, since the results from advanced refinement techniques can be used to validate the predictions from theoretical calculations. In contrast, the influence of quantum chemistry on crystallography has been more direct since the underlying refinement models are always based on theoretical calculations.[175,225,226]

One particular quantum crystallographic technique is the Hirshfeld atom refinement (HAR) approach,[227,228] that directly combines quantum chemical calculations with crystallographic refinements, and leads to more reliable E−H bond distances.[227–229] This technique has been used to obtain the main results presented in this part of the thesis.

The aim of this chapter is to give an introduction to the techniques applied in this thesis. Therefore, in Section 2.1 some relevant basics for the standard IAM refinement of X-ray structures will be summarized. This will be followed by a brief introduction to neutron diffraction in Section 2.2 because several neutron structures served as references. Finally, in Section 2.3, more advanced refinement approaches for X-ray data will be described and the focus will be particularly placed on the HAR technique.

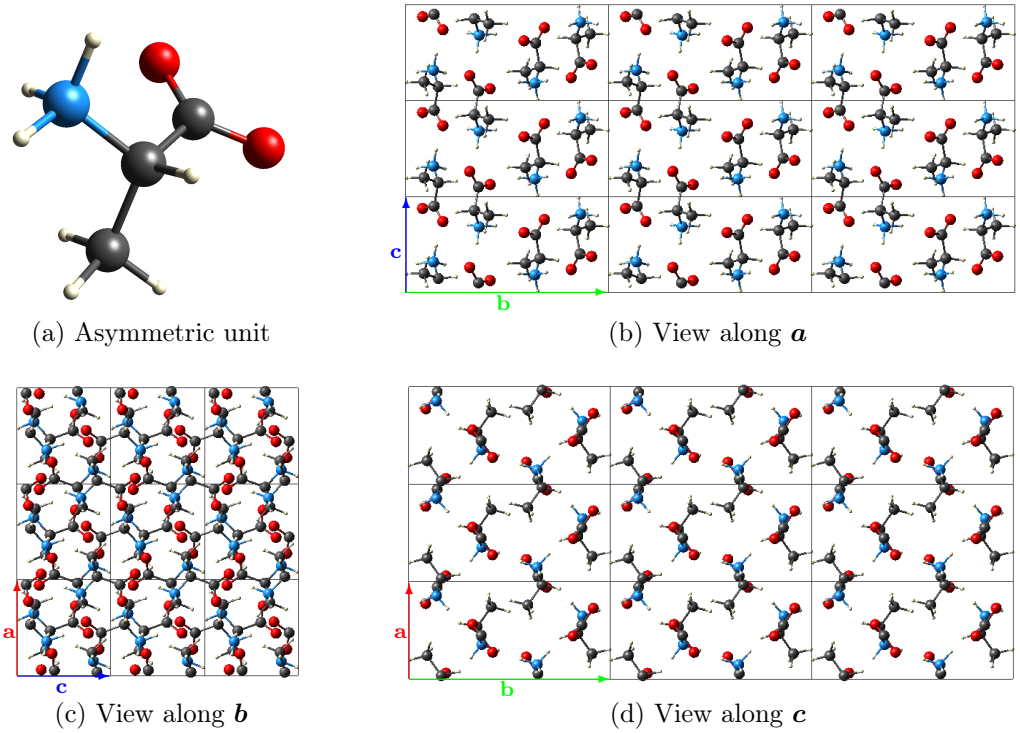
2.1 How to obtain a crystal structure

2.1.1 Description of crystals

In crystallography, a real finite crystal is usually idealized as a perfect infinite three dimensional periodic system, completely neglecting the surface of the crystal, its defects and other imperfections.[230] The idealized crystal structure can then be described by a three dimensional periodic lattice that is constructed from an infinite repetition of identical structural units that are called unit cells. The crystal lattice is characterized by three basis vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$. Since these vectors are not necessarily oriented perpendicular to each other, also the angles α (between $\boldsymbol{b}$ and $\boldsymbol{c}$), β (between $\boldsymbol{a}$ and $\boldsymbol{c}$) and γ (between $\boldsymbol{a}$ and $\boldsymbol{b}$) need to be

[1] Also other types of radiation may be used to determine electron densities, for example γ-rays and electrons.[226] However, these types of radiation are less common than X-rays[226] and will not be discussed here.

determined for the description of the lattice. The lengths of the basis vectors are the lattice constants a, b and c, which, together with the angles, characterize the unit cell that contains symmetry-equivalent units that are related to each other through the symmetry operations of the crystal. It is worth noting that the combination of all possible symmetry operations (i.e. rotation, inversion, mirroring, translation and combinations of them) leads to 230 space groups. The smallest possible combination of atoms that is sufficient to fill the complete space after applying all symmetry operations of the space group under exam is called asymmetric unit.[231] In Figure 2.2a, the asymmetric unit for the crystal structure of *L*-alanine[149] is shown. Application of the symmetry elements fills the space, which is exemplary depicted in Figures 2.2b to 2.2d for $3 \times 3 \times 3$ unit cells.



(a) Asymmetric unit (b) View along ***a***

(c) View along ***b*** (d) View along ***c***

Figure 2.2: The crystal structure of *L*-alanine:[149] (a) the asymmetric unit; (b-d) $3 \times 3 \times 3$ unit cells viewed along the different lattice axes as indicated.

2.1.2 X-ray diffraction

In 1912, Max von Laue and his co-workers discovered that X-rays are diffracted by crystals.[232] With their experiment, they did not only prove that X-rays are waves, but they also validated the description of crystals as a three dimensional periodic lattice.[211,233] They were able to explain that crystals diffract X-rays because the wavelength of X-rays is of the same order of magnitude as the distance of the lattice planes.[231,232,234] Nevertheless, they could not give a correct interpretation of the phenomenon that they had observed.[211,233] Instead, in 1913, William Lawrence Bragg and William Henry Bragg summarized the condition for constructive interference for X-rays reflected at crystal planes in their famous Bragg equation.[211,235] Their idea can be derived from Figure 2.3, where for simplicity, a single point scattering center is present at each lattice point and two incident waves are reflected by two parallel lattice planes

with spacing d. Since the lattice planes act like mirrors, the angle θ between the incoming wave and the lattice plane is equal to the corresponding angle between the reflected wave and the lattice plane. For the two waves in Figure 2.3, the condition for constructive interference is only fulfilled if the difference in the pathlength between the first and the second wave ($2l$) is equal to a multiple of the wavelength ($n\lambda$). From trigonometry it is known that, in the right triangle, the side opposed to the angle θ is given by $l = d \sin\theta$. Since the two diffracted waves will only be in phase if $2l = n\lambda$, it follows that the two waves will be in phase for any multiple n of the wavelength:

$$2d \sin\theta = n\lambda \tag{2.1}$$

which is the previously mentioned Bragg equation.[235] It follows that, for a given wavelength λ, a larger angle θ can only lead to reflections if the spacing between the lattice planes is smaller, while a smaller angle θ corresponds to a larger d-spacing.[231]

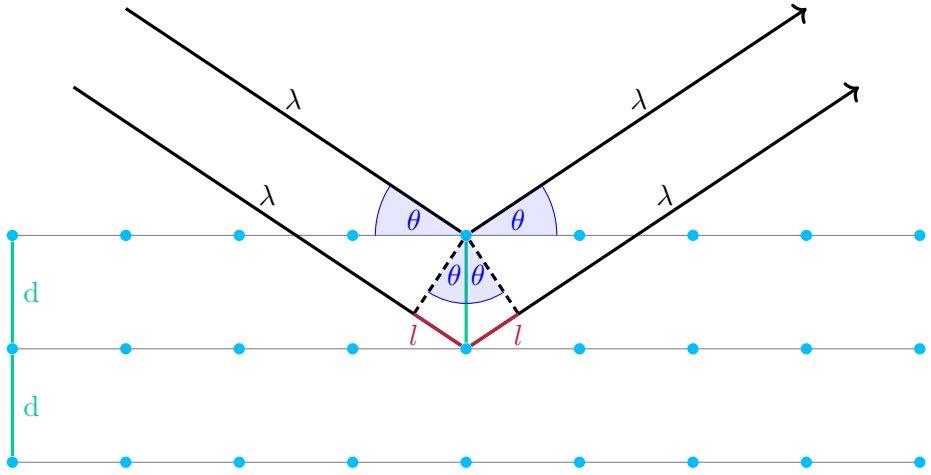


Figure 2.3: Schematic representation of Bragg's law.

The orientation of the lattice planes, where the reflections are occurring, is defined by a triad of Miller indices h, k and l. Each triad characterizes a family of lattice planes that are parallel to each other and that cut the cell axes a, b and c at the points $1/h$, $1/k$ and $1/l$. The larger the indices h, k and l are, the more often the corresponding unit cell axis is cut, and the smaller is the distance d between the parallel lattice planes. If the crystal is placed into the X-ray beam and oriented in such a way that the Bragg equation is satisfied for some of the lattice planes, one can observe a pattern of discrete diffraction spots, each associated with an index hkl, and with a characteristic position and intensity.[231]

2.1.3 From measured intensities to structure factors

In the diffraction experiment, every reflection hkl is measured with a specific intensity I_{obs} that is proportional to the amplitude of the scattered wave, which is mathematically described with the structure factor F_{obs}:

$$I_{\text{obs}}(\boldsymbol{H}) \propto |F_{\text{obs}}(\boldsymbol{H})|^2 \tag{2.2}$$

where $\boldsymbol{H}$ is related to the Miller indices h, k and l that are used to describe the lattice planes (see above). In the case of X-ray diffraction, the structure factor is the Fourier transform of

the thermally averaged electron density $\langle\rho(\boldsymbol{r})\rangle$ of the unit cell:[236,237]

$$F(\boldsymbol{H}) = \int_{\text{unit cell}} \langle\rho(\boldsymbol{r})\rangle e^{2\pi i \boldsymbol{H}\cdot\boldsymbol{r}} d\boldsymbol{r} \tag{2.3}$$

From Equation (2.3), at least in principle, the electron density could be computed from the inverse Fourier transform of the structure factor.[237,238] However, in practice this is not possible since (i) the number of reflections that can be measured in an experiment is finite, but an infinite number would be needed to compute the density; (ii) only the amplitudes of the structure factor are obtained in the experiment, while the phase information is lost (the so-called phase problem).[237] Due to these limitations, it is not possible to directly measure the electron density from an X-ray scattering experiment, instead a model is always needed.

To accomplish this task, a model structure factor is calculated from a sum of individual atomic contributions:

$$F_{\text{calc}}(\boldsymbol{H}) = \sum_{j=1}^{N} f_j(\boldsymbol{H}) e^{2\pi i \boldsymbol{H}\cdot\boldsymbol{r}_j} T_j(\boldsymbol{H}, \boldsymbol{U}_j) \tag{2.4}$$

where N is the number of atoms in the unit cell; f_j is the scattering factor of the j-th atom at position $\boldsymbol{r}_j$ and T_j is the atomic temperature factor that accounts for the motion of the atoms and depends on the atomic displacements $\boldsymbol{U}_j$ of the atom j from its average position $\boldsymbol{r}_j$.[225] In the next two subsections, the calculation of scattering factors and atomic temperature factors will be discussed.

2.1.4 The scattering factor within the independent atom model

The atomic scattering factor f_j in Equation (2.4) is a measure for the scattering power of atom j. Mathematically, it is the Fourier transform of the static electron density $\rho_j(\boldsymbol{r})$:[8]

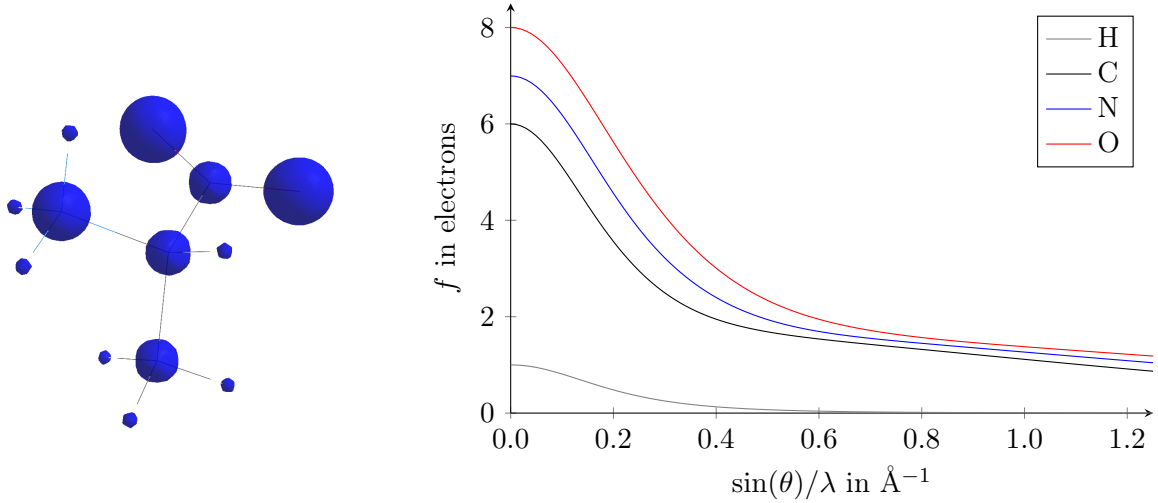
$$f_j(\boldsymbol{H}) = \int \rho_j(\boldsymbol{r}) e^{i\boldsymbol{H}\cdot\boldsymbol{r}} d\boldsymbol{r} \tag{2.5}$$

In the standard model of X-ray crystallography, the independent atom model (IAM),[220] the static atomic electron density is obtained from theoretical calculations on isolated (and often also neutral) atoms and spherically averaged afterwards.[8] The sum of these atomic densities is also called the promolecular density (compare Figure 2.4a). For the computation of the scattering factors for most neutral atoms, relativistic HF calculations are used.[225,239]

The scattering factor does not only depend on the individual scattering power of each atom, but, for a given wavelength, also on the scattering angle θ. The latter can be explained from the fact that, contrary to the simplified picture of the point scattering centers used for the explanation of the Bragg equation in Section 2.1.2, the size of a real atom is similar to the wavelength of the X-rays. Therefore, only those parts of the atomic density that lay directly on the lattice planes contribute strictly in phase. The scattering of the other parts is subject to a certain phase shift that increases with $\sin\theta/\lambda$. Due to this phase shift, the scattering factors f become smaller for larger values of $\sin\theta/\lambda$.[231]

In summary, the scattering factors are normalized to the number of electrons for each element and their values decrease as $\sin\theta/\lambda$ becomes higher. The values of the scattering factors are stored in tables (for example in the International Tables of Crystallography, reference

[239]) or internally in crystallographic programs. Scattering factors for selected elements are shown in Figure 2.4b. It is worth noting that the exact dependence of the scattering factors on the angle θ varies for the different elements according to the "form" of the corresponding electron density. Therefore, the scattering factors are also called atomic form factors. The rapid decrease of the scattering factors at higher angles θ leads on average to reflections with weaker intensity. This is especially severe for hydrogen atoms.[231]



(a) Promolecular density

(b) Scattering factors

Figure 2.4: (a) Promolecular density for *L*-alanine, isovalue=0.5 e/au^3. (b) Atomic scattering factors for the elements in *L*-alanine. Values calculated from equation 6.1.1.15 and Table 6.1.1.4 in reference [239].

2.1.5 Atomic displacement parameters

A further approximation has been made so far, namely that the atoms in the crystal are located at fixed positions. However, in the real crystal this is not the case. In fact, atoms vibrate around their mean positions. These vibrations lead to an additional phase shift. In analogy to what has been discussed in the previous subsection, the phase shift due to the vibration increases the more the atoms are vibrating and the larger the scattering angle θ becomes. Hence, the scattering power of each atom is further decreased with its vibrations.[231]

The motion of the atoms in the crystal is described using atomic displacement parameters. In the simplest case, the atomic vibrations are assumed to have the same amplitude in all directions. Within this approximation, the motion of the atom can be described using one isotropic displacement parameter. In reality, an atom vibrates with different amplitudes in the different directions. For this more complicated case, six individual anisotropic displacement parameters (ADPs) $U_{ij} = U_{11}, U_{22}, U_{33}, U_{12}, U_{13}, U_{23}$ are needed to describe the vibrations of each atom. These parameters can be used to define a thermal ellipsoid that represents the space in which the center of the atomic electron density can be found with a probability of usually 50%.[231,240] For structures of small molecules based on the IAM, in most cases, ADPs are used for non-hydrogen atoms, while isotropic displacement parameters are used for hydrogen atoms.

2.1.6 From structure factors to the refined crystal structure

The model structure factor can be computed using Equation (2.4) by exploiting the tabulated IAM scattering factors, the positions of the atoms and their atomic displacement parameters. However, the atomic positions are initially unknown. This is because only intensities are obtained from the diffraction experiment, while the information about their phases is lost. If the phases of the reflections were known, the electron density could be computed from structure factors, by using the following expression:

$$\rho(\boldsymbol{r}) = \frac{1}{V}\sum_{hkl} F(\boldsymbol{H})e^{-2\pi i \boldsymbol{H}\cdot\boldsymbol{r}} = \frac{1}{V}\sum_{hkl} |F(\boldsymbol{H})|e^{i\phi(\boldsymbol{H})}e^{-2\pi i \boldsymbol{H}\cdot\boldsymbol{r}} \tag{2.6}$$

where V is the volume of the unit cell, $|F(\boldsymbol{H})|$ is the structure factor amplitude and $\phi(\boldsymbol{H})$ the corresponding phase. The positions of all non-hydrogen atoms correspond to maxima in the electron density. Therefore, if the electron density was known, also the positions of the non-hydrogen atoms could be obtained. The obstacle of the lost phases is called the "phase problem", which needs to be solved to get a structure.[231,241]

To accomplish this task, different strategies are available. All of them make some prior assumptions about the molecular structure or the electron density to obtain information on the phases. For solving crystal structures of small molecules, direct methods are the most common ones. They are based on the facts that the electron density in the crystal can only take positive values and that scattering takes place on discrete atoms. Then, by exploiting statistical relationships between the intensities of the reflections, direct methods can be used to finally estimate the phases of the reflections.[231,241]

In this way, an initial model for the structure is obtained. However, this model is not in optimal agreement with the experimental data. Therefore, the model needs to be further improved (for example by correcting the atom types and adding missing atoms) and refined by optimizing at least the positional and atomic displacement parameters and an overall scale factor for the structure. Hence, a minimum of nine parameters need to be refined for each non-hydrogen atom (three positional parameters and six ADPs), while only four parameters are needed for each hydrogen atom (three positional parameters and one for the isotropic displacement parameter). Additionally, other parameters can be included, for example occupancy factors of the atoms. For a reliable and stable refinement a minimum number of observed reflections per refined parameter is required and a data-to-parameter ratio of eight to ten is usually recommended.[242]

In practice, the refinement is done exploiting a least-square procedure that minimizes the mean-square difference between $|F_{\text{obs}}|$ and $|F_{\text{calc}}|$:

$$Q = \sum_{hkl} w(|F_{\text{obs}}| - |F_{\text{calc}}|)^2 \tag{2.7}$$

$$Q = \sum_{hkl} w(F_{\text{obs}}^2 - F_{\text{calc}}^2)^2 \tag{2.8}$$

where the factor w is a weight that takes into account that not all the data has been measured with the same accuracy.[231] If Equation (2.7) is used, the refinement is said to be "based on

F_{obs}", which was the standard procedure until the 1990s. However, since for very weak reflections F^2_{obs} can be negative, its square root cannot be calculated. One way to avoid this problem is to suppress the weak data (e.g. $F^2_{obs} < 2\sigma(F^2_{obs})$) in the refinement, with the disadvantage that valuable information is lost. Another way to avoid both problems is to use Equation (2.8), so that the refinements are carried out against F^2_{obs} data, where all reflections can be used in the refinement and no square roots of negative numbers need to be computed. This refinement based on F^2_{obs} is nowadays the usual one.[231]

2.1.7 Validation of the refinement quality using R values

To judge the quality of the resulting model, the final agreement between the observed and calculated structure factors is usually given by the R value

$$R = \frac{\sum_{hkl} ||F_{obs}| - |F_{calc}||}{\sum_{hkl} |F_{obs}|} \tag{2.9}$$

or the wR_2 value:

$$wR_2 = \frac{\sum_{hkl} w(F^2_{obs} - F^2_{calc})^2}{\sum_{hkl} w(F^2_{obs})^2} \tag{2.10}$$

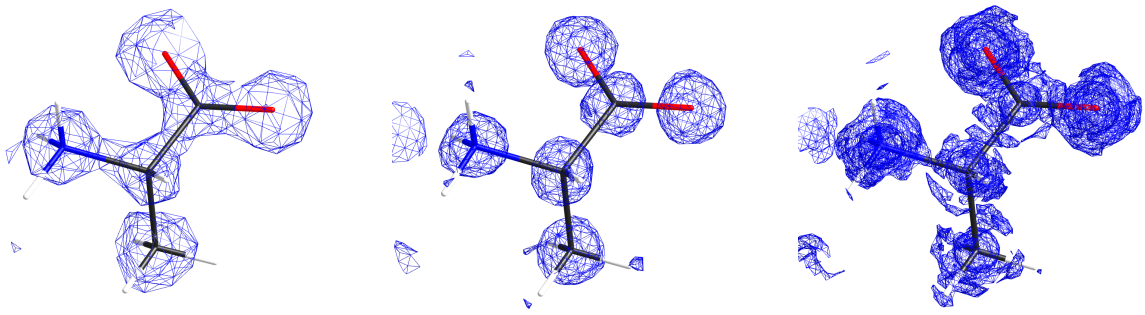
In general, the lower these two values are, the better is the agreement between the observed and calculated structure factors.[231] R values and other quantities that measure the statistical agreement between the structure factors are also called figures of merit.

2.1.8 Electron density maps and influence of the resolution

The primary product of the structure solution and refinement procedures outlined in Section 2.1.6 is a model for the electron density.[243] Electron densities are usually displayed using different types of maps. For example, $|F_{obs}|$ values can be used together with the estimated phases to generate the so called F_{obs} maps.[231,243] In Figure 2.5, examples for these maps are shown in dependence of the resolution for a dataset of L-alanine[244] that has a very high maximum resolution of 0.45 Å. The resolution of a dataset is defined as the minimum lattice plane spacing d that corresponds to the highest angle θ at which reflections have been measured.[243]

For Figures 2.5a and 2.5b, the resolution of the L-alanine dataset has been cut to values of 1.2 Å and 0.8 Å, respectively, while for Figure 2.5c the complete dataset[244] has been used. The resolution of 1.2 Å (Figure 2.5a) is also known as the atomic resolution limit.[245] For d-spacings larger than 1.2 Å, the electron density maxima overlap, hence the individual atomic contributions cannot be separated for solving a structure with direct methods.[245–247] To obtain a sufficient data-to-parameter ratio, the resolution of a small molecular dataset should be at least between 0.84 Å and 0.77 Å (compare Figure 2.5b).[242,248] However, it is important to note that for a high-quality measurement, it is necessary to measure both high and low resolution reflections. This is because the low resolution data make a large contribution to the electron density maps, while the high resolution reflections provide finer details of the map,[247] as can be seen in Figure 2.5c.

Another important type of map is the residual density map that represents an additional measure for the agreement between observed and calculated data. It is based on computing the



(a) 1.2 Å (b) 0.8 Å (c) 0.45 Å

Figure 2.5: Dependence of the F_{obs} maps on the resolution. The figure shows the structure of *L*-alanine[244] with F_{obs} maps plotted using the 1.5 e/Å^3 isosurface for the following resolution cutoffs: (a) 1.2 Å, (b) 0.8 Å and (c) 0.45 Å.

difference between the observed and calculated structure factor amplitudes using calculated phases:[249]

$$\rho_0(\boldsymbol{r}) = \frac{1}{V}\sum_{hkl}(|F_{\text{obs}}| - |F_{\text{calc}}|)e^{i\phi_{\text{calc}}}e^{-2\pi i\boldsymbol{H}\cdot\boldsymbol{r}} \tag{2.11}$$

Residual density maps visualize areas, where positive values indicate structural features that were measured but not modeled and *vice versa*. Therefore, the maps can be used for building a structure model, for example hydrogen atoms are often added to the initial structure model from the maximum peaks in the residual density.[231,243] Additionally, since the residual density contains all sources of errors that are related to the model, the data processing, experimental errors or noise, it can give valuable insights for improving the description of electron densities in crystals.[249] For example, residual density maps can be used to judge the ability of a model to describe the measured data. In Figure 2.6, the residual density after a refinement with the IAM is shown, where the green areas correspond to features in the density that are not described by the IAM and that are typically associated with the deformation of the electron density due to chemical bonding. Therefore, from Figure 2.6, it can be concluded that the high-resolution measurement of *L*-alanine[244] contains information that cannot be modeled using the promolecular density of the IAM.

2.1.9 Refinements with restrictions

Under certain circumstances, for example if the resolution of the measurement and the data-to-parameter ratio are low, a free refinement of positional and atomic displacement parameters can be difficult or impossible. In these cases, restrictions are usually introduced to the refinement. In this context, crystallographers differentiate between constrained or restrained refinements. Constraints can be used to fix certain parameters. For example, groups of atoms with well known geometries (phenyl rings, etc.) can be refined as rigid groups with fixed bond lengths and angles. Therefore, constraints reduce the number of parameters in the refinement. In contrast, in the restrained refinements all atoms are refined, but some expected values for certain parameters are included. For example, an aromatic ring can be restrained to be planar. From a mathematical point of view, restraints are added as additional observations to

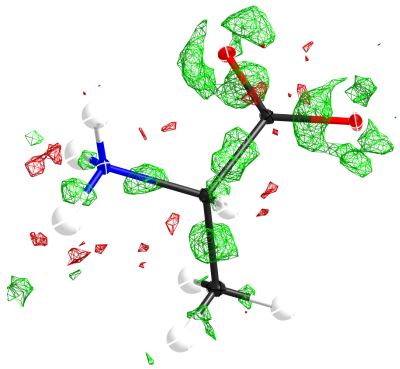


Figure 2.6: Residual density after IAM refinement of the *L*-alanine structure.[244] Positive residual densities are shown in green, negative ones in red. For both of them the 0.15 e/Å^3 isosurface was used.

the refinement.[231,242]

For refinements of small molecules, restraints and constraints are particularly important for disordered structures.[242] In general, there are two main types of disorder: static disorder and dynamic disorder. In the former, the types or orientation of individual atoms or fragments can differ. In the latter, some atoms or fragments are moving, for example methyl groups can rotate. In the refinement, disorder is usually described for the main configurations of the atoms that are assigned different occupancy values.[231]

For the refinement of large molecules, in particular proteins, the resolution and the data-to-parameter ratio is usually too low for a free refinement. However, since the geometries of the amino acids are well known, constraints and restraints should be used.[231]

2.1.10 Refinement of hydrogen atoms

While exact positions of non-hydrogen atoms can be obtained reliably and routinely using the standard refinement procedure described above, this is not the case for hydrogen atoms. In fact, refinements based on the IAM cannot yield accurate positions for hydrogen atoms. This is because these atoms possess only one electron that is forming a bond with another atom. Therefore, the electron density maximum for the hydrogen atoms is shifted inside the bond leading to a deformation of the electron density. This cannot be captured by the spherical description used in the IAM. Hence, the corresponding E−H bond lengths are always too short.

In general, there are different ways to obtain positions for the hydrogen atoms in the refinement:[250]

- The positions can be freely refined using the IAM. The corresponding bond lengths are systematically shifted inside the bonds, leading to E−H bond lengths that are on average 0.12 Å shorter than the correct ones.[229,251]
- The positions of the hydrogen atoms can be calculated using expected values according to geometrical criteria. In the following constrained refinement (see previous subsection), they are fixed using rigid groups or riding models, where the movement of the hydrogen atom is coupled with the one of its bonding partner.[231]

- Refinements of neutron diffraction data reliably provide accurate positions of hydrogen atoms.[231,252,253] If neutron structures are available for the compound under exam, the E−H bond lengths can be constrained to these values. However, in most of the cases, the neutron data for the investigated compound are not available. Alternatively, the bond lengths can be normalized to average neutron values, for example those from Bruno and Allen.[254] More details about neutron diffraction will be given in Section 2.2.
- Since theoretical calculations that take into account the crystal environment can lead to accurate hydrogen positions,[255,256] they could be performed for the structure under exam and the corresponding E−H bond lengths could be used as constraints in the refinement.
- Finally, the IAM can be replaced by more advanced refinement strategies that explicitly take into account the aspherical electron density of the atoms.[227,228,251,255,257–262] One option is to use the Hirshfeld atom refinement (HAR)[227,228] that was shown to give accurate bond lengths for organic molecular compounds[229] and recently also for some compounds with hydrogen atoms bonded directly to heavy elements.[262] More details about HAR and other techniques exploiting aspherical scattering factors will be given in Section 2.3.

Obtaining atomic displacement parameters for hydrogen atoms is generally even more challenging.[263] Also here, different possibilities exist and some examples will be given below.

- In most standard refinements, the motion of hydrogen atoms is simply described by isotropic displacement parameters.[231] However, this is a crude approximation that severely affects the static electron density in covalent bonds.[263]
- If neutron data are available for the same compound, the neutron ADPs can be used in the refinements of X-ray data.[252,264–266] However, since the access to neutron diffraction experiments is limited (compare Section 2.2), this option is usually not available.
- Alternatively, the *SHADE* (simple hydrogen anisotropic displacement estimator) program[267,268] can be used. The corresponding hydrogen ADPs are derived from high-quality neutron structures of organic compounds, from spectroscopic or neutron diffraction experiments for which the information is provided by the user, or from periodic *ab initio* calculations. Other options to estimate hydrogen ADPs from combined experimental and theoretical data exist as well. For example, one may use an approach based on ONIOM calculations,[269] the *APD-Toolkit*,[270] which is based on the optimized geometries and frequency analyses of the molecules in the invariom database[271] (for details about the database, see Section 2.3.2) or *NoMoRe*,[272,273] which exploits periodic *ab initio* calculations.
- In principle, ADPs for hydrogen atoms can also be obtained from refinements based on aspherical models,[258,274,275] for example from HAR.[227,228] However, the HAR ADPs are usually less accurate than those obtained with the *SHADE* server and in some cases even non-positive definite (NPD). Therefore, HAR should not be considered as a standard procedure for obtaining ADPs.[255,275,276] Nevertheless, it has been shown that accurate E−H bond lengths can be obtained with HAR, even when isotropic displacement parameters are used for hydrogen atoms.[276]

2.2 Neutron diffraction

While X-rays interact with the electrons in an atom, neutrons interact with the nuclei.[236] The underlying formalism of neutron diffraction is very similar to the one of X-ray diffraction, except that the atomic scattering factors used for X-ray diffraction need to be replaced by neutron scattering lengths b. In analogy to the model structure factor for X-ray scattering (compare Equation (2.4)), a corresponding factor for neutron scattering can be defined as:

$$F_{\mathrm{calc}}(\boldsymbol{H}) = \sum_{j=1}^{N} \bar{b}_j e^{2\pi i \boldsymbol{H}\cdot\boldsymbol{r}_j} T_j(\boldsymbol{H}, \boldsymbol{U}_j) \tag{2.12}$$

where $\bar{b}_j$ is the averaged nuclear scattering length of atom j.[236] However, while X-ray scattering factors can be obtained from theoretical calculations, neutron scattering lengths cannot be predicted from nuclear theory and are thus obtained experimentally.[236,277]

Compared to the wavelength of neutrons ($\lambda \approx 1$ Å), the nuclei are very small (≈ 0.00001 Å) and act as point scatteres. Therefore, contrary to the previously described X-ray scattering amplitudes, the corresponding neutron scattering lengths are independent of $\sin\theta/\lambda$.[236] Additionally, the neutron scattering length is independent of the neutron wavelength for most nuclei.[239] Examples for the coherent scattering lengths of different elements are shown in Figure 2.7.

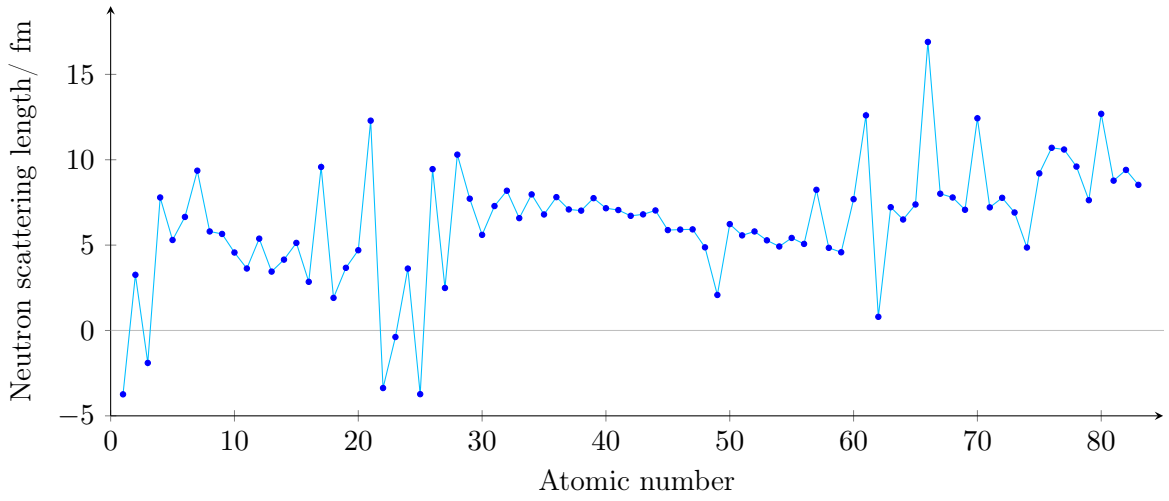


Figure 2.7: Coherent neutron scattering lengths for the elements from hydrogen to bismuth in their natural abundance. Values as given in Table 4.4.4.1. of reference [278].

As can be seen in Figure 2.7, the scattering lengths of neighboring atoms in the periodic table differ. Additionally, each isotope of an element has a characteristic scattering length.[239] Therefore, neutron diffraction experiments have the advantage that neighboring elements and isotopes can be easily distinguished,[279] while this is not possible for X-ray diffraction experiments. Furthermore, another advantage of neutron diffraction is that information on the nuclear positions can be obtained directly, while they are obtained indirectly in the X-ray experiment from the electron density distribution.[236] Since hydrogen is one of the few elements with a negative scattering length (compare Figure 2.7), it can be easily distinguished from other elements, because it appears in the negative nuclear density maps.[253] Therefore,

neutron diffraction experiments currently represent the most reliable experimental technique to determine the positions and ADPs of hydrogen atoms. For this reason, the results of neutron refinements are, in general[229,257–259,266] and throughout this thesis, used as references.

However, neutron diffraction experiments also have disadvantages. In particular, they can only be carried out at nuclear reactors or spallation sources. The access to these large facilities is usually quite limited, and their operation is accompanied by non-negligible costs (and risks).[229,279] Furthermore, there are also practical limitations since the crystals for neutron experiments need to be significantly larger than for X-ray diffraction. In particular, a crystal size of 1 mm^3 is typically needed, which is difficult to grow for most compounds.[238] Therefore, it is desirable to establish other methods that are accompanied by lower risks and costs and that are easier to access.[229]

2.3 Towards more accurate crystal structure refinements using aspherical scattering factors

The standard refinement strategy outlined in Section 2.1 is based on the independent atom model. This model is very successful because it is universally applicable and the refinements are straightforward. However, it was also mentioned that for good quality and high-resolution data, there is a significant amount of measured data that cannot be modeled by the spherical description associated with the IAM (compare Figure 2.6). This unmodeled density corresponds to the deformation of the electron distribution due to chemical bonding. To account for this deformation, it is necessary to use an aspherical description of the electron density in the models underlying the refinements.

2.3.1 The multipole model

During the 1970s, several approaches were proposed for modeling the aspherical electron density, where the one that has become the most popular is the Hansen and Coppens multipole model.[237,280,281] In this approach, the total electron density is divided into atomic contributions (the so-called pseudoatoms) that are the smallest transferable atomic electron density fragments from which the total electron density can be reconstructed.[259] The electron density for each pseudoatom is further divided into three parts: a spherical core electron density, a spherical valence density and an aspherical deformation density. While the core electron density remains frozen during the refinement, the valence shell is allowed to expand or contract and to take an aspherical form. The electron density for each pseudoatom can be Fourier transformed to obtain the corresponding aspherical atomic scattering factor that is used in the refinement.[237,259]

In 1975, Stewart and coworkers proposed an approach[282] for a more accurate determination of hydrogen positions based on generalized scattering factors that were obtained from finite multipole expansions on diatomic molecules. This approach has been successfully applied for multipole model refinements of different small molecules, yielding element-hydrogen bond lengths that were significantly elongated compared to the IAM values.[244,263,283–285] Nevertheless, the applicability of the multipole model is limited to datasets of very high resolution (at least 0.5 Å)[259,286] and the number of parameters that need to be refined using the

multipole model is significantly higher compared to an IAM refinement.

2.3.2 The transferable aspherical atom model

In order to also allow the refinement of data with less exceptional quality (e.g., for large systems as proteins), databases of multipole parameters have been constructed.[251,271,287,288] They are based on the observation that the multipole parameters are transferable between different molecules if the atoms have similar chemical environments.[289] The stored parameters have been obtained either from high-resolution X-ray experiments (ELMAM[287] database) or from theoretical densities (invariom[271] and UBDB database[288]). In the corresponding transferable aspherical atom model (TAAM), the multipole parameters are transferred from the database to the system under exam and are kept fixed during the refinement.[251] Applying TAAM, the electron density can be reconstructed for small molecules and also for macromolecules.[290–292] For the latter, certain requirements should be met concerning the resolution, the atomic displacement parameters and the amount of disorder and solvent in the macromolecular structure.[293] Several studies showed that TAAM refinements of small molecular crystal structures yield element-hydrogen bond lengths that are comparable to the corresponding neutron or theoretically obtained reference values.[251,255,257–259]

2.3.3 The bond-oriented deformation density model

A simpler but less sophisticated approach has been recently implemented in the refinement software *SHELXL*,[294] which is the most widely used program for the refinement of small molecule crystal structures. In this new approach,[260] it is possible to obtain aspherical scattering factors that take into account the deformation in the electron density associated with bonding and lone pair regions by placing Gaussian functions in appropriate positions. The corresponding parameters are obtained from DFT calculations, and are not refined freely. The underlying model is called bond-oriented deformation density (BODD). It has been reported that the new formalism can be used to obtain meaningful positions of hydrogen atoms.[260]

2.3.4 The Hirshfeld atom refinement

An alternative to the multipole models and to tabulated scattering factors is represented by the Hirshfeld atom refinement (HAR) technique that was introduced in 2008 by Jayatilaka and Dittrich[227] and further developed by Capelli *et al.* in 2014.[228] The resulting HAR procedure is illustrated in Figure 2.8.

The starting point for HAR is usually an IAM structure and the corresponding experimental structure factors. Since 2014, HAR is based on an iterative procedure that consists in the following steps:

1. A single-point quantum mechanical calculation is performed on a molecular system that corresponds at least to the asymmetric unit of the crystal structure under exam. Henceforth, this system will be also called "molecular reference unit".
2. The electron density of the molecular reference unit is partitioned into contributions for each atom using the Hirshfeld stockholder partitioning method.[295,296] In this technique,

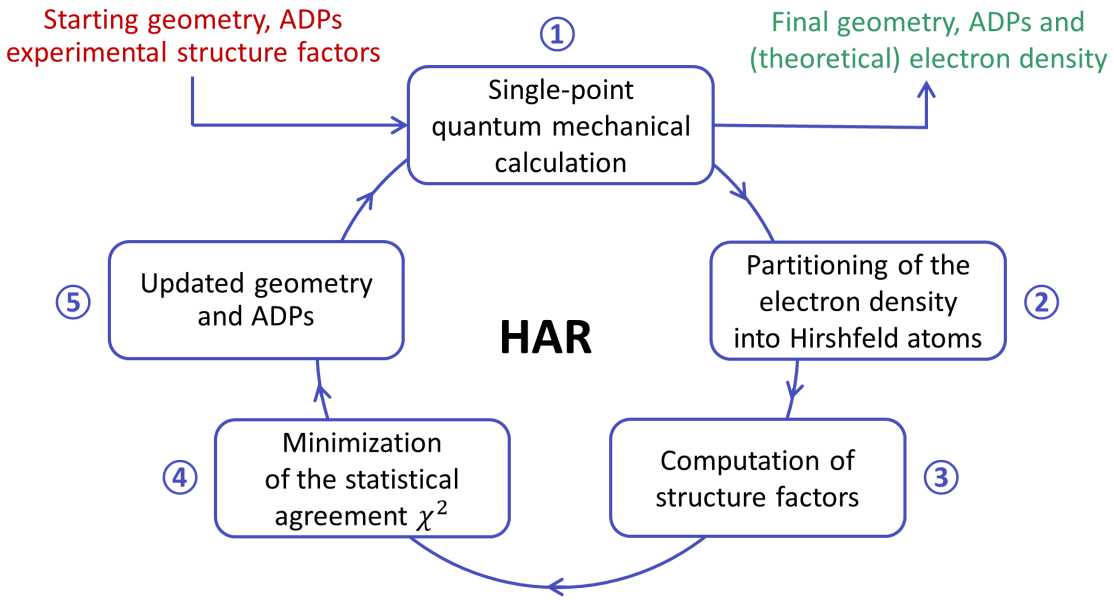


Figure 2.8: Schematic representation of the Hirshfeld atom refinement procedure. Adapted with permission from [221]. Copyright 2020 Elsevier B.V.

the static atomic electron density $\rho_A(\boldsymbol{r})$ for a generic atom A is given by

$$\rho_A(\boldsymbol{r}) = w_A(\boldsymbol{r})\rho(\boldsymbol{r}) \tag{2.13}$$

where $\rho(\boldsymbol{r})$ is the electron density resulting from the previous single-point calculation and $w_A(\boldsymbol{r})$ is the Hirshfeld partitioning function for atom A. This function is defined by

$$w_A(\boldsymbol{r}) = \frac{\rho_A^0(\boldsymbol{r})}{\sum_{B=1}^{N_{\text{atoms}}} \rho_B^0(\boldsymbol{r})} \tag{2.14}$$

where $\rho_A^0(\boldsymbol{r})$ is the spherical density of atom A in the promolecule, which is divided by the promolecular density for the complete molecular reference unit to obtain $w_A(\boldsymbol{r})$.

3. The aspherical atomic density functions obtained from the previous step are Fourier transformed to obtain the corresponding atomic scattering factors. These are afterwards used in Equation (2.4) to compute the structure factors.
4. The structure is refined adopting a least-square procedure to minimize the statistical agreement factor χ^2:

$$\chi^2 = \frac{1}{N_r - N_p} \sum_{hkl} \frac{(\eta|F_{\text{calc}}| - |F_{\text{obs}}|)^2}{(\sigma)^2} \tag{2.15}$$

where N_r is the number of reflections, N_p the number of parameters in the model, σ the experimental uncertainty for the observed structure factor amplitude $|F_{\text{obs}}|$ and η an overall scale factor that places the calculated structure factor amplitudes $|F_{\text{calc}}|$ on the same scale of the experimental ones.

5. Convergence is checked for the refinement parameters (atomic coordinates and ADPs). If convergence is reached, the refinement is terminated and a final model for the structure

and, optionally, a corresponding wavefunction are obtained. Otherwise, if convergence is not reached, the procedure re-starts from point 1 with a single-point calculation on the updated geometry.

From the procedure described above, important differences between HAR and refinements based on the multipole model and the IAM arise. While the IAM is based on tabulated spherical atomic scattering factors, the aspherical scattering factors in HAR are tailor-made for the structure and newly calculated at each iteration of the procedure.[8] Furthermore, the number of refined parameters per atom is significantly larger in refinements based on the multipole model by Hansen and Coppens compared to IAM refinements, whereas HARs and IAM refinements require the same number of parameters per atom.[227]

Statistical validation of the HAR technique for more than 80 small molecular organic compounds showed that the resulting E−H bond lengths were on average from 0.01 to 0.02 Å shorter than the corresponding neutron reference values, while the IAM bond lengths were on average approximately 0.12 Å shorter.[229] Very recently, it has also been shown that HAR can yield accurate E−H bond lengths, where E is an heavy element.[262] Moreover, a direct comparison of HAR with refinements based on the Hansen and Coppens multipole model, TAAM and IAM showed that the aspherical refinement techniques yield E−H bond lengths that are in general in good agreement with the neutron values, but HAR gave the most accurate results among the different aspherical models.[259]

HAR was originally implemented in the software *TONTO*.[297] A minimal version of HAR, called HARt, has been included in the crystallographic package *olex2*.[298,299] More recently, a graphical interface called *lamaGOET*[300] has been developed in order to allow user-friendly preparations of HARs with all possible functionalities. Furthermore, *lamaGOET* offers the possibility to use other quantum chemical software for single-point calculations (compare step 1 in the HAR procedure as described above and in Figure 2.8). In particular, *lamaGOET* provides an interface to *Gaussian09*,[150] *Orca*[301,302] or *ELMOdb*[105] on one side and to *TONTO* on the other side. This gives access to a full variety of QM methods for the first step of the HARs procedure. The idea at the basis of *lamaGOET* has been further developed giving rise to the implementation of *NoSpherA2* in *olex2*.[303] *NoSpherA2* is an interface that allows the use of any of the above mentioned programs for single-point calculations, and that subsequently performs the Hirshfeld partitioning, computes the corresponding aspherical atomic scattering factors, and passes them to *olex2.refine*.[304] Hirshfeld atom refinements with *NoSpherA2* are completely independent of the original software *TONTO* and many of the previous limitations are overcome. For example, thanks to *NoSpherA2*, HARs using restraints and constraints and HARs of disordered structures are now possible.[303] Another advantage is that the refinements are no longer carried out against F values (as in *TONTO*) but against F^2 (compare Section 2.1.6).

Nevertheless, when the work for this thesis started in 2018, there was still (and currently also is) some room to improve HARs. In particular, the following challenges for HAR have been addressed and will be discussed in the next chapters:

- Due to the fact that HARs require repetitive fully QM calculations, they are computationally more expensive than refinements based on the IAM, TAAM or BODD models.

- Due to the cost of the underlying QM calculations, HAR had never been applied to refine macromolecules, although, as mentioned in the beginning of this chapter, X-ray diffraction is the main source for obtaining macromolecular structures.
- In *TONTO*, the only implemented QM methods are HF and DFT with the BLYP[305,306] and B3LYP[306–309] functionals. For this reason, post-HF methods had never been applied in HARs. The development of *lamaGOET* opened the door for testing the performance of other QM methods in the framework of HARs.
- From the very first implementation of HAR, a cluster of point charges and dipoles, usually called cluster charges, was used in the QM calculation to simulate the crystal environment of the molecular reference unit.[227] It has been shown that their use significantly improves the agreement with the corresponding neutron values for the bond lengths involving hydrogen atoms.[227,228,298] However, more advanced, fully QM embedding techniques had never been tested.
- Finally, the exact description of heavy atoms remains challenging for HAR. In particular, accurate and precise positions for hydrogen atoms that are bonded directly to a heavy element can be obtained only in some cases.[262,303]

In the following four chapters, all the above described challenges will be addressed in some way. In all cases, the first step of the HAR procedure was varied. In Chapter 3, the results of a coupling between ELMO libraries with HAR will be presented. With the corresponding HAR-ELMO technique[149] structures of small molecules can be refined at a significantly lower computational cost. It also allows the refinement of macromolecular crystal structures. Moreover, in Chapter 4, the performance of different QM methods for HARs will be compared. In particular, it will be evaluated whether post-HF methods are necessary for HARs. Furthermore, in Chapter 5, the QM/ELMO embedding technique will be employed to introduce a fully QM description of the surrounding crystal environment into the first step of the HAR procedure. Finally, in Chapter 6, a combination of HAR with the QM/ELMO technique will be proposed. This novel HAR-QM/ELMO strategy could be useful for the refinement of compounds that include a hydrogen atom that is bonded directly to a heavy element and offers the possibility to further evaluate the necessity for post-HF methods for HARs.

3 Hirshfeld atom refinements for large systems: the HAR-ELMO method

The fundamental difference between HAR and IAM refinements is that in the former the aspherical atomic scattering factors are tailor-made for the investigated system and updated at every iteration of the HAR procedure (compare Figure 2.8), while in the latter the spherical scattering factors are stored in tables for each atom type. Therefore, due to the underlying quantum chemical calculations, which need to be repeated for every iteration, the computational cost of HAR is significantly higher than for refinements based on the IAM. Moreover, as discussed in Section 1.1.4, already one QM calculation on large systems (such as proteins) is very challenging, and performing it iteratively is impracticable to impossible. Nevertheless, exploiting the main advantage of HAR, namely that accurate E−H bond lengths can be obtained, would be desirable not only for small molecular structures but also for macromolecules.[310] In fact, approximately half of the atoms in a protein are hydrogen atoms[247,311] and obtaining accurate and precise hydrogen positions in these macromolecules is fundamental for understanding their functions and reactivity.[247,310,312]

However, before HAR can be applied to proteins, it is necessary to speed up the first step of the corresponding procedure (Figure 2.8) by using cheaper methods of quantum chemistry. Several such strategies have been introduced in Chapter 1. In principle, any technique that includes a fully QM description of the molecular reference unit and provides an electron density could be used to speed up the HAR procedure. For example, the MFCC method (compare Section 1.2) has recently been combined with HAR giving rise to the *fragHAR* technique that was successfully applied for the refinement of different oligopeptide systems.[88] This chapter is dedicated to an alternative approach, namely the HAR-ELMO refinement technique, consisting in the coupling of the ELMO libraries (see Section 1.3.4) with HAR.[149]

To validate the reliability of the HAR-ELMO approach, it was applied to an amino acid and a dipeptide. The obtained structures were compared to those resulting from traditional HARs based on HF calculations and to IAM and neutron structures. Furthermore, HAR-ELMO was applied to refine the structures of two polypeptides and a protein.[149] All these refinements were performed by Dr. Lorraine Andrade Malaspina (University of Bremen / University of Bern, group of Dr. Simon Grabowsky). Nevertheless, since the starting idea of performing this work came from our group, I was also involved in this study, which represented the starting point of the work presented in this part of the thesis. Therefore, in this chapter, the HAR-ELMO technique and the most important results from reference [149] will be summarized. In Section 3.1, the adopted procedure for HAR-ELMO refinements will be outlined and in Section 3.2, some selected results obtained in the validation of the technique will be described. In Section 3.3 and Section 3.4, the refinements of two polypeptides and a small protein will be briefly discussed, respectively. In the fifth and last section of this chapter, conclusions will be drawn, and in addition to the original study, the possibility of using the HAR-ELMO strategy for the routine refinements of macromolecules will be critically evalu-

ated.

3.1 The HAR-ELMO technique

The HAR-ELMO technique is based on alternating steps of (i) transfers of ELMOs and computation of the electron density; and (ii) Hirshfeld partitioning of the obtained density, computation of the structure factors and least-square refinement of the structure. In summary, the ELMO transfer replaces the single-point calculation in step 1 of the HAR procedure (compare Section 2.3.4 and Figure 2.8), while steps 2-5 of the procedure remain unaltered. The transfer of ELMOs and the computation of the corresponding electron density are performed by the *ELMOdb* program[105] (compare Section 1.3). The remaining steps of the HAR procedure are carried out by the software *TONTO*.[297] To automatize the alternating transfer of updated electron densities and new structures between the two programs, and to ensure that both steps are repeated until convergence of the structure is reached, the *lamaGOET* interface[300] has been developed.

The purpose of the study published in reference [149] was to test the HAR-ELMO technique, by validating it against traditional HAR, IAM and neutron refinements and by applying it to biomolecules and organometallic compounds. In this chapter, only the results for the former class of compounds will be summarized.

For the HAR-ELMO refinements, ELMOs pre-computed with the basis set 6-311G(d,p) were transferred from the ELMO libraries at each HAR iteration. The corresponding traditional HARs were performed at HF/6-311G(d,p) level. All refinements were set up using the *lamaGOET* interface.

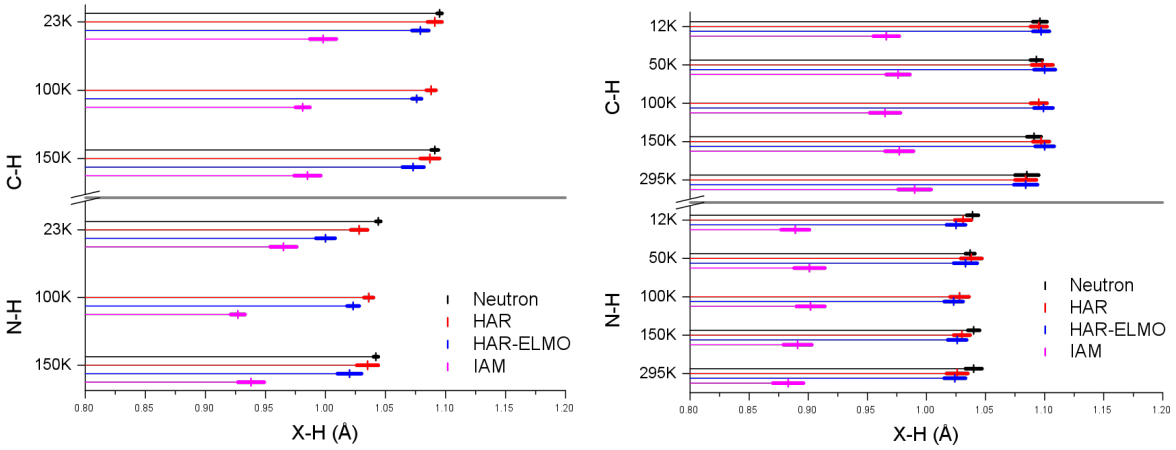
3.2 Validation on *L*-alanine and glycyl-*L*-alanine

To validate the HAR-ELMO strategy, the structures of the amino acid *L*-alanine and of the dipeptide glycyl-*L*-alanine were refined with the HAR-ELMO technique, traditional HARs and IAM refinements. For *L*-alanine, X-ray datasets with three different temperatures (23 K, 100 K and 150 K) were used. The 23 K dataset was collected by Destro and co-workers,[244] while the other two were measured for this study.[149] For comparison, neutron datasets were also collected at 23 K and 150 K.[149] For glycyl-*L*-alanine, five different X-ray datasets obtained at 12 K, 50 K, 100 K, 150 K and 295 K were taken from the literature.[228] Corresponding neutron datasets were also available except for the 100 K data.[313]

Focusing first on the comparison of the E−H bond lengths, the average N−H and C−H bond lengths resulting from the refinements of *L*-alanine and glycyl-*L*-alanine for all the different temperatures, are shown in Figure 3.1. In the structures of glycyl-*L*-alanine, the average bond lengths obtained with HAR-ELMO agree with the HAR and neutron results within one sample standard deviation. In contrast, the average E−H bond lengths resulting from IAM refinements are significantly too short. The HAR-ELMO average bond lengths agree with the HAR and neutron ones independently of the temperature. However, the agreement is better for the C−H bonds than for the N−H ones, for which the average HAR bond lengths are slightly shorter than the neutron ones. Furthermore, the average HAR-ELMO bond lengths

are slightly shorter than the HAR ones. Similar trends can be found for the E−H bond lengths in *L*-alanine, especially for the N−H ones obtained for the 23 K dataset. A better agreement with the neutron results could probably be achieved if the influence of the crystal environment was treated explicitly in the computations. For example, a surrounding cluster of charges could be included in the calculations to mimic the environment and to polarize the electron density of the molecular reference unit.[227,228,261,298] This option was not considered here because the polarization of ELMOs is currently not implemented. Nevertheless, for almost all types of E−H bond lengths in the two systems, the ELMO approximation does not reduce the accuracy of the HAR results.

Moreover, the computational cost for obtaining these results is much lower for HAR-ELMO than for traditional HARs, as can be seen in Table 3.1 for the refinements of the *L*-alanine and glycyl-*L*-alanine 150 K datasets. For these two datasets, Table 3.1 also lists the R values and the χ^2 values. The R values resulting from IAM refinements are significantly higher than the ones for HAR-ELMO that are closer to those from traditional HARs. This can be expected for good quality datasets, where the aspherical description should lead to a better agreement between the calculated and the measured structure factors. Both the R and χ^2 values in Table 3.1 are higher for HAR-ELMO than for HAR, which can be ascribed to the approximations made in the computation and transfer of ELMOs.



(a) L-Ala (b) Glycyl-*L*-alanine

Figure 3.1: Average E−H bond lengths with corresponding standard deviations upon averaging in (a) *L*-alanine and (b) glycyl-*L*-alanine obtained from neutron, traditional HAR, HAR-ELMO and IAM refinements. Adapted with permission from reference [149]. Copyright 2019 American Chemical Society.

3.3 Application to polypeptides

After its successful validation, the HAR-ELMO refinement technique was applied to two polypeptides, namely Leu-enkephalin[314] (entry GEWWAG01 in the Cambridge structural database (CSD)) and a fibril-forming segment of the human prion protein[315] (PDB code 2OL9). Both X-ray datasets were collected at 100 K. However, while the resolution of the former is very high (0.43 Å), the one of the latter is significantly lower (0.85 Å), but still high in comparison to the majority of structures in the PDB (see Section 3.5 below).

Table 3.1: CPU wall-clock timing (format dd:hh:mm:ss) and refinement statistics for HAR and HAR-ELMO refinements. The values in brackets are the R values resulting from IAM refinements. Adapted with permission from reference [149]. Copyright 2019 American Chemical Society.

Structure	Method	CPU wall time	R1(F)	χ^2
Gly-L-Ala (150K)	HAR	00:00:12:57	0.0161	2.336
	HAR-ELMO	00:00:03:36	0.0168	2.725
			(0.0251)	
L-Ala (150K)	HAR	00:00:05:55	0.0202	3.152
	HAR-ELMO	00:00:02:25	0.0210	3.523
			(0.0279)	
Leu-enkephalin	HAR	00:09:52:00	0.0422	0.505
	HAR-ELMO	00:01:44:17	0.0430	0.545
			(0.0557)	
Fibril-forming segment	HAR	01:07:00:00	0.0436	9.903
	HAR-ELMO	00:00:22:56	0.0474	11.913
			(0.0446)	
Crambin (d=0.54Å)	HAR	Impracticable	-	-
	HAR-ELMO	09:23:47:53	0.0715	5.004
			(0.0704)	
Crambin (d=0.73Å)	HAR	Impracticable	-	-
	HAR-ELMO	06:00:15:16	0.0624	7.672
			(0.0618)	

Concerning the refinement of the crystal structure of Leu-enkephalin, the original deposited structure already included all atoms that were subsequently refined freely using the IAM, HAR and HAR-ELMO strategies. The vibrations of the non-hydrogen atoms were described with ADPs, whereas isotropic displacement parameters were used for all the hydrogen atoms.

In contrast, the original structure of the fibril-forming segment contained five ordered water molecules but no hydrogen atoms. Prior to the preliminary IAM refinement, hydrogen atoms were added to the structure of the polypeptide according to geometrical criteria, whereas the hydrogen atoms of the water molecules and one hydrogen atom in the hydroxyl group of serine (atom H1D) were added using the *tleap* software of the *AMBER* molecular dynamics package[148]. Most of the atoms in the fibril-forming segment were refined freely with the IAM, HAR and HAR-ELMO strategies, except for the atom H1D and the water molecules, whose positional and displacement parameters were kept fixed during the refinements. This treatment was necessary because the original software for performing HARs, *TONTO*, does not allow refinements with restraints and *NoSpherA2* (see Section 2.3.4) had not been developed at the time of the study.

The average E−H bond lengths obtained from the IAM, HAR-ELMO and traditional HAR refinements of the structures of Leu-enkephalin and the fibril-forming segment are shown in Figure 3.2. In lack of neutron data for the two polypeptides, the E−H bond lengths are compared to average neutron E−H bond lengths for organic molecules collected by Allen and Bruno.[254] In most of the cases, the average E−H bond lengths in the two polypeptide structures are significantly elongated for the traditional HAR and HAR-ELMO refinements

compared to the IAM results. Moreover, for the HAR-ELMO refinements, all of the C−H and O−H bond lengths in both polypeptides, and the N−H bond lengths in the fibril-forming segment agree with the average neutron values within a single sample deviation. Only the mean N−H bond lengths in Leu-enkephalin are shorter than the average neutron values. This result could probably improve using the same strategy that was already mentioned in the previous subsection, namely by explicitly considering the crystal environment in the first step of the HAR procedure.[227,228,298]

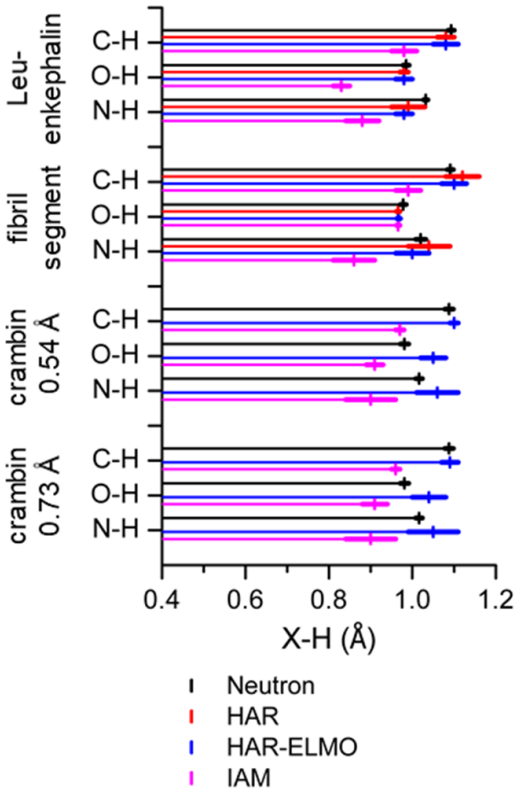
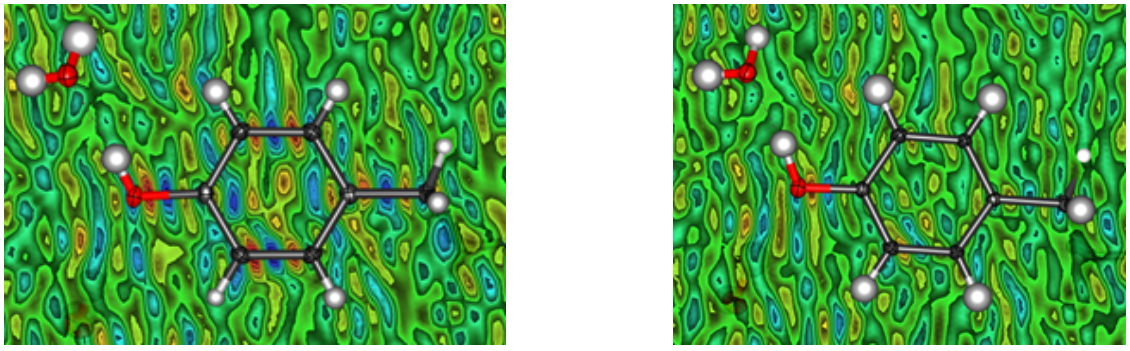


Figure 3.2: Average E−H bond lengths with corresponding standard deviations upon averaging in the structures of Leu-enkephalin, the fibril-forming segment and crambin obtained from IAM, HAR-ELMO and, when possible, traditional HAR refinements. For comparison, average values from neutron structures are also shown.[254] Adapted with permission from reference [149]. Copyright 2019 American Chemical Society.

The two different resolutions of the polypeptide datasets have significant influence on the R values (compare Table 3.1) and the residual densities. For Leu-enkephalin a significant reduction can be observed for the R value that takes a value of 0.0557 for the IAM refinement and drops to values of 0.0422 and 0.0430 for the HAR and HAR-ELMO refinements, respectively. This can be explained from the residual density maps shown for the tyrosine group in Figure 3.3. In Figure 3.3a, significant regions (red and blue contour lines) of unmodeled residual density are located in the areas of chemical bonds. These regions vanish after the refinement with the HAR-ELMO technique, indicating that the dataset contained information that could not be described with the spherical densities in the IAM but needed an aspherical description.

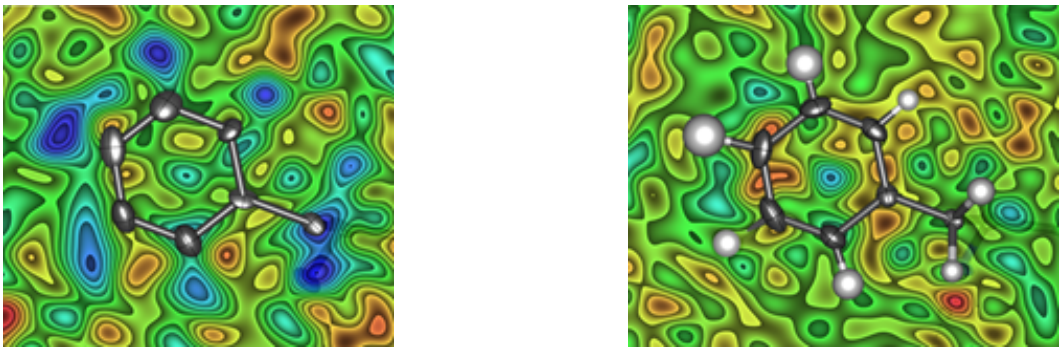
In contrast, for the fibril-forming segment the traditional HAR and HAR-ELMO refine-



(a) IAM (b) HAR-ELMO

Figure 3.3: Residual density map for Leu-enkephalin in the plane of the tyrosine group for (a) the IAM and (b) HAR-ELMO refinements. Contour minimum and maximum values: -0.2 eÅ^{-3} and 0.3 eÅ^{-3}, respectively; contour interval: 0.05 eÅ^{-3} interval. Color code: from blue (positive), through green (zero) to red (negative). ADPs depicted at 50% probability level.

ments do not improve the R values compared to the IAM. Figure 3.4 shows the residual density maps in the plane of the phenyl group without hydrogen atoms as deposited in the PDB and with hydrogen atoms after HAR-ELMO refinement. Although the residual density maps for the deposited structure (compare Figure 3.4a) clearly indicate the missing hydrogen atoms (blue areas), the dataset does not include information about the deformation of the valence density. Moreover, the quality of the dataset is not good enough for free refinement of all atoms, leading to unphysical[255] NPD atoms for the IAM, HAR and HAR-ELMO refinements. Nevertheless, the HAR-ELMO refinement provides significantly elongated C−H bond lengths, which are on average 1.07(14) Å long for the the aromatic ring depicted in Figure 3.4. In contrast, the corresponding bond lengths obtained from the IAM refinement are on average 0.96(8) Å, which is significantly shorter than the average neutron value[254] of 1.083(17) Å for aromatic C−H bond lengths.



(a) (b)

Figure 3.4: Residual density map for the fibril-forming segment in the plane of the phenyl group (a) without hydrogen atoms as deposited in the PDB and (b) with hydrogen atoms after HAR-ELMO refinement. Contour minimum and maximum values: -0.2 eÅ^{-3} and 0.3 eÅ^{-3}, respectively; contour interval: 0.05 eÅ^{-3} interval. Color code: from blue (positive), through green (zero) to red (negative). ADPs depicted at 50% probability level.

3.4 Application to a protein

Finally, the HAR-ELMO technique was also applied to refine the structure of a small protein. More precisely, the high-resolution X-ray data for the protein crambin (PDB code 1EJG) was taken from the literature.[316] Since the structure contains an amount of disordered solvent water molecules and since the refinement of disorder is not implemented in the software *TONTO*, the *SQUEEZE* routine[317] was used to calculate the scattering contributions from the solvent molecules and exclude them from the original experimental structure factors. Furthermore, for the disordered parts of the protein, only those atoms belonging to the major conformation were kept in the refinement and full occupancies were assigned to these atoms.

The refinement of crambin was tested for two different resolutions, namely the maximum resolution of the dataset of 0.54 Å and a reduced one with a resolution of 0.74 Å. Using both datasets, a preliminary unconstrained IAM refinement was performed with the software *TONTO*. Based on the obtained structure, HAR-ELMO and another IAM refinement were performed. In both refinements, the positional and displacement parameters were freely refined for 436 of the total 642 atoms in the structure of crambin. The remaining atoms were fixed using the parameters of the previous unconstrained IAM structure. The fixed atoms either belong to the disordered regions of the protein structure or were fixed to guarantee convergence of the refinement. In all three refinements of the crambin structure, isotropic displacement parameters were used for the hydrogen atoms, while anisotropic ones were used for the non-hydrogen atoms.

For the freely refined atoms in the HAR-ELMO and IAM structures, the E−H bond lengths were extracted and averaged. The results are shown in Figure 3.2. As for the polypeptides, the bond lengths are compared to average neutron values from Allen and Bruno.[254] For both resolutions, all the E−H bond lengths in the HAR-ELMO structures are significantly elongated compared to the corresponding IAM ones. For the structures obtained with the HAR-ELMO strategy, the average C−H bond lengths agree well with the neutron reference, while the O−H and N−H bond lengths are longer than the corresponding neutron ones. A possible explanation for this could be that the HAR-ELMO refinement indeed overestimates the bond lengths. However, the values from Allen and Bruno[254] are an average of different bond lengths found in structures of small molecules and may not be the optimal reference in this situation. In particular, Allen and Bruno excluded long O−H and N−H bond lengths from their average values because they observed that the corresponding hydrogen atoms were involved in intramolecular hydrogen bonds that were characterized by short distances between the hydrogen and acceptor atoms.[254] Since intramolecular hydrogen bonds are frequent in protein structures one could speculate that these bond lengths are missing in the Allen and Bruno reference values.

Like for the fibril forming segment, the R value for the HAR-ELMO structure of crambin does not improve compared to the IAM (compare Table 3.1), which is probably related to the missing information in the dataset despite the high resolution. As mentioned above, traditional HARs are impracticable for proteins because they would require a new QM calculation of the complete protein for every HAR iteration. In contrast, the cheaper transfer of ELMOs allows Hirshfeld atom refinements of proteins. The total computational cost of a refinement

using the HAR-ELMO strategy depends on the size of the molecule, or more precisely on the number of atomic orbitals, and on the number of reflections in the least-square refinement. For crambin, 96 139 reflections are included in the refinement of the higher resolution dataset, which is more than the double of the number of reflections for the lower resolution one (45 265 reflections). Therefore, the refinements with HAR-ELMO took almost 10 days for the higher resolution dataset, and reduced to 6 days for the one with lower resolution (see Table 3.1).

Different properties of the protein can be computed from the wavefunction obtained after the last refinement iteration, for example, different bonding descriptors such as the electron localizability indicator (ELI-D),[318] the electrostatic potential (ESP) and NCI[319,320] or IGM[321,322] plots (for more details about these methods see Part II). In Figure 3.5, the deformation density, the ESP and the ELI-D are shown for the complete crambin molecule. They were computed using the software *cuQCT* that allows the rapid calculation of these properties on graphics cards.[149,323]

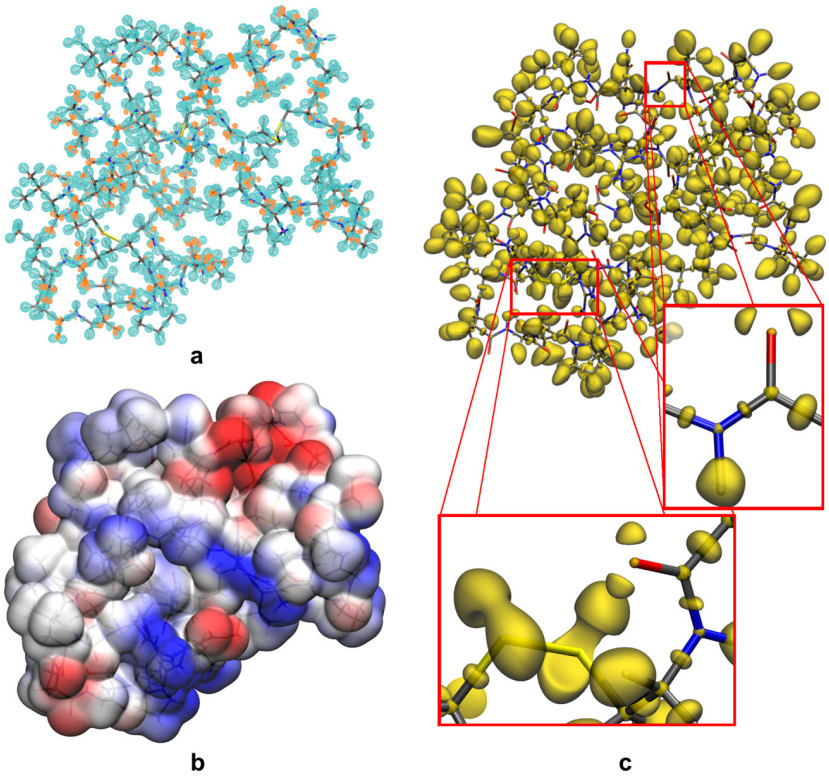


Figure 3.5: Molecular properties for crambin calculated on the final HAR-ELMO geometry and wavefunction: (a) isosurfaces of the deformation density, positive isovalue 0.04 eÅ^{-3} with blue color, negative isovalue -0.04 eÅ^{-3} with red color; (b) electrostatic potential in the range from -0.1 to +0.1 e bohr^{-1} plotted on the electron density isosurface with an isovalue of 0.001 e bohr^{-1}; (c) ELI-D isosurfaces with an isovalue of 2.2 for the main figure and 2.0 for the magnified pictures; all for the entire protein crambin. Reprinted with permission from reference [149]. Copyright 2019 American Chemical Society.

3.5 Conclusions and outlooks

In this chapter, the new HAR-ELMO refinement strategy has been described and the corresponding results obtained for biomolecules have been summarized. In general, it can be concluded that the HAR-ELMO strategy allows for refining small systems with similar results as obtained for the traditional HAR technique but at a significantly reduced computational cost. Therefore, HAR-ELMO refinements are still feasible for large systems such as proteins, where traditional HARs are impracticable or impossible.[149]
Do the above-presented results indicate that the HAR-ELMO strategy could be applied routinely in protein crystallography as expected by Cachau *et al.*?[310] Before this goal could be accomplished in the future, there are two main limitations that need to be adressed, namely the general quality of protein datasets and the lack of disorder refinement in the software *TONTO*.

Beginning with the former, HAR requires datasets with a resolution of approximately 0.8 Å.[229] However, the majority of protein crystal structures either do not reflect or are not measured up to this resolution. This is shown in the left plot of Figure 3.6, where the number of X-ray structures in the PDB is depicted for different resolution ranges. However, the plot on the right side of Figure 3.6 shows that the number of structures with subatomic resolution is steadily growing,[310,311] also for the resolutions below 0.8 Å. Moreover, despite measurements with subatomic resolution (below 1 Å) are still rare for proteins, their importance should not be underestimated because they provide (i) more accurate structures that can be obtained with less prior assumptions; (ii) more details about the solvent molecules; and (iii) the possibility to correctly identify and refine the positions of hydrogen atoms.[310]

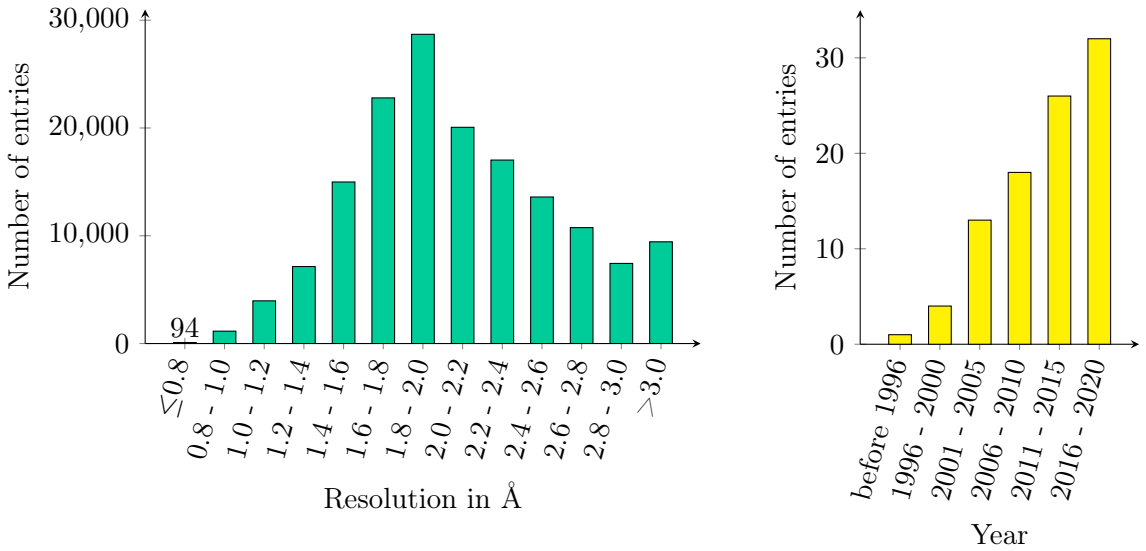


Figure 3.6: The left plot shows the number of X-ray structures in the protein data bank for a given resolution range. The right plot gives the number of entries with a resolution lower or equal to 0.8 Å depending on the year of deposition. Data from `www.rcsb.org`, accessed on June 7, 2021.

For the third point it is of fundamental importance to also address the second limitation mentioned above, namely the missing possibility to refine disorder. In fact, the refinement of the protein crambin was only possible using quite severe approximations, where a substantial part of the structure was kept fixed during the refinement. This could be avoided by using a different refinement software. In fact, as mentioned in Section 2.3.4, the new development of *NoSpherA2*[303] in *olex2* allows for refining structures using restraints and constraints, so that the severe approximations made in this study could be replaced by strategies that are routinely used for the refinement of proteins.[243] We have recently implemented the HAR-ELMO refinement strategy inside *NoSpherA2* and, in collaboration with the Grabowsky group, we are currently performing a new refinement of crambin using the utilities of the *olex2.refine* software.

In addition to the crucial implementation of disorder in the refinement, further improvements of the HAR-ELMO technique can be also envisaged. For example, the transfer of ELMOs could become significantly faster by developing a parallelized version of the *ELMOdb* program. Furthermore, HAR could be coupled with the QM/ELMO embedding technique giving rise to the HAR-QM/ELMO approach that would allow the treatment for example of metal-centers in proteins at quantum mechanical level. Although an application of the HAR-QM/ELMO technique to a protein will not be presented in this thesis, a variant of the method for including the effects of the crystal environment in structure refinements of small molecules will be described in Chapter 5 and the application of the HAR-QM/ELMO technique to the refinement of metal-organic compounds will be introduced in Chapter 6.

In summary, before HAR-ELMO can become a routine technique for the refinement of protein structures, further improvements are needed both from the software development point of view as well as from the experimental side. However, if the necessary steps are pursued, the HAR-ELMO technique could provide better structures with more accurate E−H bond lengths also for proteins.

4 Hirshfeld atom refinements based on post-Hartree-Fock methods

4.1 Introduction

As already mentioned in the previous chapters, the crucial step for every HAR is the underlying single-point calculation. Since the only program that could perform HARs used to be *TONTO*, the available QM methods were limited to those implemented in that software. In particular, these were HF and DFT with two functionals (BLYP and, more recently, also B3LYP).[227,324] Additionally, relativistic variants were implemented in *TONTO* and the influence of relativistic effects on the structure factors was evaluated.[325–327] Moreover, HAR was implemented in the software *VASP* to perform periodic *ab initio* calculations.[328]

Capelli *et al.* observed that the choice of the QM method has a systematic effect on the bond lengths involving hydrogen atoms that result from HAR. In their HARs of the glycine-*L*-alanine structure, the E−H bond lengths obtained with HF are systematically too long, while those obtained with BLYP are systematically too short compared to the neutron references. They speculated that a hybrid DFT functional like B3LYP could provide better bond lengths.[228]

Thanks to the development of the *lamaGOET* interface, it became possible to also use other software to obtain the electron densities for HARs. This opened the door for systematically testing different QM methods. For example, the performance of HAR with more advanced QM techniques, in particular post-HF ones, had never been evaluated. However, as discussed in Section 1.1.2, these methods are based on a multi-determinant wavefunction *ansatz* and, at least in principle, should provide better electron densities than the HF strategy.

Therefore, we decided to investigate whether the post-HF densities could improve the structures resulting from HARs. In particular, we tested the performances of MP2 and CCSD methods for HAR, and compared the obtained results to those resulting from HARs with underlying HF and DFT calculations.[221] Since the ELMO method is central to the work presented in this thesis, in this chapter also HAR-ELMO refinements will be used for comparison. Most of the results and discussion described in this section were previously published in reference [221]. However, the results obtained from the HAR-ELMO refinements were added specifically for this thesis.

4.2 Computational details

4.2.1 X-ray and neutron data

All refinements in this study were performed using the X-ray dataset of *L*-alanine collected by Destro and coworkers.[244] This dataset was chosen for the following reasons: (i) *L*-alanine is a reasonably small molecule, hence computationally expensive wavefunction techniques

such as coupled cluster are feasible (for a discussion about the computational cost associated with wavefunction calculations see Section 1.1.4); (ii) the X-ray dataset of *L*-alanine is of high quality and high resolution (0.46 Å) and is often used for benchmarking new refinement strategies; (iii) structural parameters obtained from the refinement of neutron data[149] collected the same low temperature (23 K) are available as references. For all refinements in this study, only reflections for which $F_{\text{obs}} > 4\sigma(F_{\text{obs}})$ were used.

4.2.2 Hirshfeld atom refinements

The following HARs were performed in this study: HAR-ELMO, HAR-HF, HAR-BLYP, HAR-B3LYP, HAR-MP2 and HAR-CCSD refinements. All the single-point calculations were carried out in combination with three different basis sets that were chosen according to a ranking proposed by Fugel *et al.*,[298] who classified the basis sets def2-SVP as adequate, def2-TZVP as excellent and def2-TZVPP as benchmark quality for HARs. The ELMOs for *L*-alanine were computed on the geometry corresponding to the neutron structure and adopting the same three basis sets. For comparison, we also performed an IAM refinement using the software *TONTO*.

4.2.3 The *lamaGOET* interface

All HARs were carried out using the interface *lamaGOET*.[300] Starting from an IAM structure, the interface allowed us to automatically perform the usual iterative steps; with (i) single-point calculations using *Gaussian09*[150] or transfers of ELMOs using *ELMOdb*[105] (point 1 of the HAR procedure outlined in Section 2.3.4 and Figure 2.8); and (ii) structure refinements using *TONTO* (points 2-5 of the HAR procedure). All these steps were repeated until convergence of the structure was reached.

4.3 Results and discussion

To evaluate the influence of the different QM methods and basis sets on the results obtained from HARs, we focused on differences in: (i) the structure factor based quantities, such as crystallographic figures of merit as well as deformation and residual densities; (ii) the lengths of the O−H and N−H bonds in *L*-alanine; (iii) the ADPs; and (iv) the electron densities obtained from the different single-point calculations or ELMO transfers.

4.3.1 Structure factor based descriptors

At the beginning of this subsection, we will focus on discrepancies in the structure factors based descriptors obtained from all the different HARs and the IAM refinement. Afterwards, the differences between the individual HARs will be discussed in detail.

As mentioned in Section 2.1.7, it is common practice to report the agreement between measured and calculated structure factors in terms of *R* values that are lower when the agreement is better. For HARs, the final χ^2 value (see Equation (2.15)) is also usually given. An optimal agreement between the structure factors would result in a χ^2 value of 1, meaning that the modeled structure factors are on average within one standard deviation of

the experimental values. However, in practice χ^2 often converges to higher values. Therefore, the value of χ^2 should be as close as possible to the optimal value of 1. The χ^2 and R values resulting from all the different HARs in this study are given in Table 4.1. From this table, it can be observed that both the χ^2 and R values decrease significantly when HARs are performed instead of IAM refinements. For example, the value of χ^2 is 5.260 after the IAM refinement, while it lies between 1.296 and 1.868 for all HARs. Analogous trends can be observed also for the R values.

Table 4.1: χ^2 and R values for the Hirshfeld atom and independent atom model refinements performed on *L*-alanine. Adapted with permission from reference [221]. Copyright 2020 Elsevier B.V.

χ^2	def2-SVP	def2-TZVP	def2-TZVPP
MP2	1.421	1.331	1.336
CCSD	1.418	1.296	1.302
HF	1.630	1.425	1.437
ELMO	1.868	1.523	1.537
BLYP	1.364	1.411	1.414
B3LYP	1.362	1.318	1.323
IAM	5.260		

$R(F)/\%$	def2-SVP	def2-TZVP	def2-TZVPP
MP2	1.922	1.882	1.887
CCSD	1.919	1.864	1.868
HF	2.007	1.921	1.925
ELMO	2.085	1.966	1.971
BLYP	1.891	1.903	1.906
B3LYP	1.892	1.871	1.875
IAM	2.798		

$R_w(F)/\%$	def2-SVP	def2-TZVP	def2-TZVPP
MP2	1.631	1.579	1.582
CCSD	1.629	1.558	1.562
HF	1.747	1.634	1.640
ELMO	1.870	1.689	1.697
BLYP	1.598	1.626	1.627
B3LYP	1.597	1.571	1.574
IAM	3.163		

The reason for the lower figures of merit resulting from HAR compared to IAM lies in the fact that the high-resolution and high-quality dataset of *L*-alanine contains more information than the simple spherical IAM can describe. As mentioned in Section 2.1.8 in the description of Figure 2.6, the regions where the electron density is poorly modeled correspond to chemical bonds and lone pairs. Since the QM wavefunctions underlying the different HARs and the aspherical Hirshfeld partitionings account for the deformation of the electron density due to chemical bonding, the residual densities resulting from HARs should be statistically distributed in the unit cell and the unmodeled regions associable with chemical bonds and lone pairs should disappear. In Figure 4.1, the residual densities are shown in the plane of the carboxylate group of *L*-alanine for the different refinements. In fact, all maps corresponding

to HARs show significantly less features than the one for the IAM, although also for HARs the residuals are not completely randomly distributed.

In addition to the residual density maps, the fractal distribution of the residual density can be accessed through so-called Meindl-Henn plots[249] that, in the ideal case, should follow a narrow parabolic distribution, with a balance between positive and negative values and a maximum at null residual density. In Figure 4.2, the Meindl-Henn plots resulting from all our refinements of *L*-alanine are shown. They were obtained using the *jnk2RDA* software.[249] The previously described parabolic trend is indeed obtained for all plots associated with HARs except for the one based on the transfer of ELMOs with basis set def-SVP, which shows a shoulder in the negative region. Nevertheless, this plot is still significantly narrower than the one associated with the IAM refinement that also shows some shoulders, especially in the negative region.

On the basis of the quantities considered above, the effects of the basis set choice on the HARs will be now discussed. For all HARs except for the one based on calculations with the BLYP functional, *R*- and χ^2 values decrease when passing from the split-valence basis set def2-SVP to the triple-zeta basis sets def2-TZVP and def2-TZVPP (compare Table 4.1). For example, comparing the values obtained with the basis sets def-SVP and def-TZVP, the differences in χ^2 amount to -0.090, -0.122, -0.205, -0.047 and -0.044 for MP2, CCSD, HF, BLYP and B3LYP, respectively. The values of χ^2 are more similar for the two triple-zeta basis sets, although with def2-TZVPP slightly larger values are always obtained, with differences in χ^2 between +0.003 to +0.012. Analogous trends are also observed for the R values (compare again Table 4.1).

The analyses of the residual density maps (Figure 4.1) and of the Meindl-Henn plots (Figure 4.2) further confirm that also the residual densities resulting from HARs based on calculations using the two triple-zeta basis sets are highly similar. In fact, the residual density maps include more features for refinements carried out with def2-SVP than for the refinements with the triple-zeta basis sets. Also the Meindl-Henn plots are broader for the smallest basis set than for the larger ones. In contrast, the differences between the residual density maps and between the Meindl-Henn plots resulting from HARs with the triple-zeta basis sets are very small.

Therefore, it is possible to conclude that, for the examined structure-factor based quantities, refinements carried out with the triple-zeta basis sets generally lead to higher agreements between calculated and observed structure factors than the ones with the double-zeta basis set def2-SVP. The small differences observed between the two triple-zeta basis sets probably arise from the fact that, in the present study, the only difference between the two triple-zeta basis sets is the number and type of polarization functions for the hydrogen atoms, while for all the other element types present in *L*-alanine the two basis sets are identical. This is probably the reason why the residual density maps and the Meindl-Henn plots are very similar for these two basis sets. Furthermore, since the R and χ^2 values are lower for def2-TZVP than for def2-TZVPP, it can be concluded that the extra polarization functions on the hydrogen atoms do not improve the agreement with the experimental data because the experiment is not sensitive to the polarization of the hydrogen atoms. Furthermore, it should be noted that, in general, a larger basis set does not automatically lead to better results when comparing

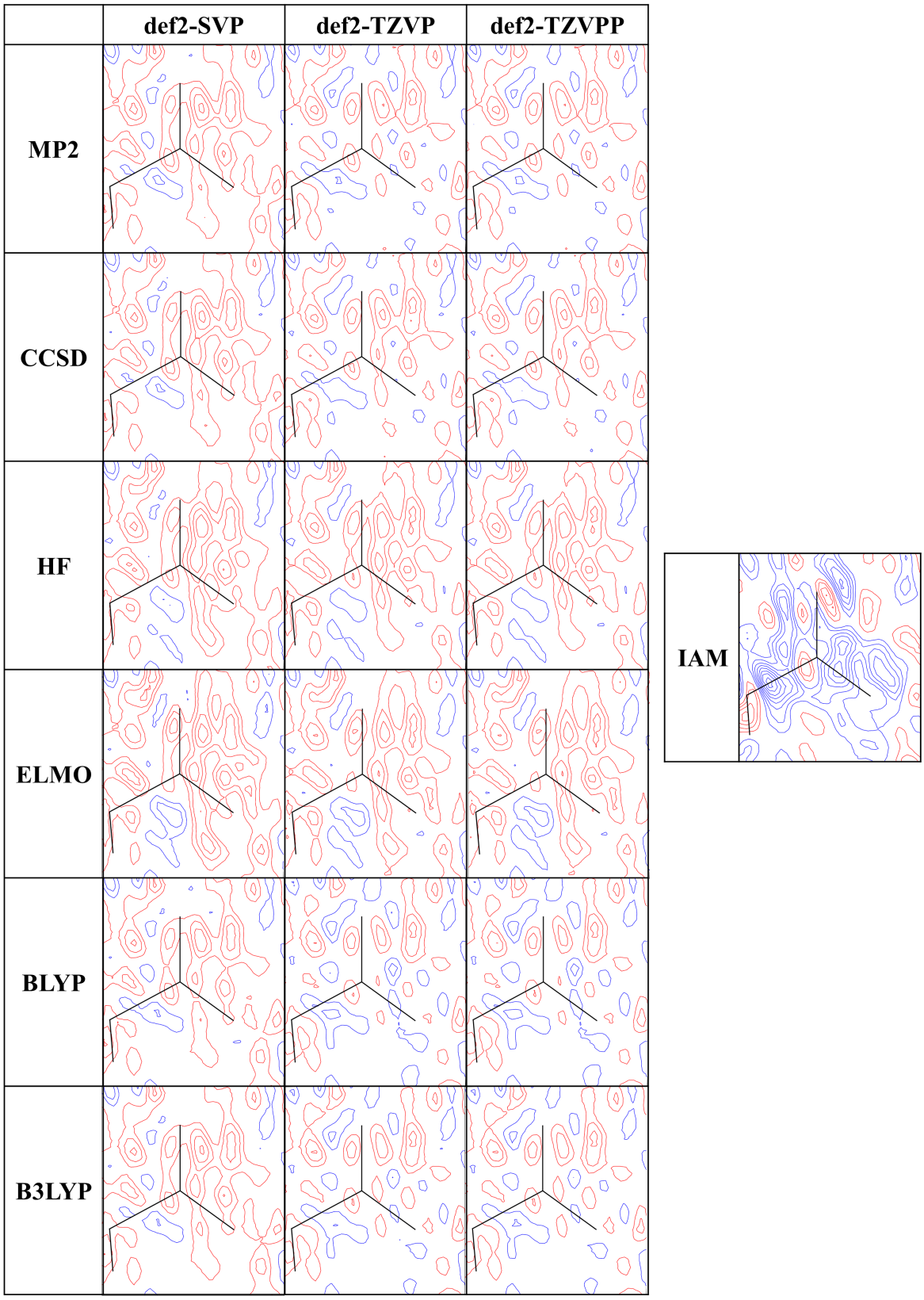


Figure 4.1: HAR and IAM residual densities in the plane of the carboxylate group of *L*-alanine. Contour level: 0.02 eÅ^{-3}; blue = positive; red = negative.

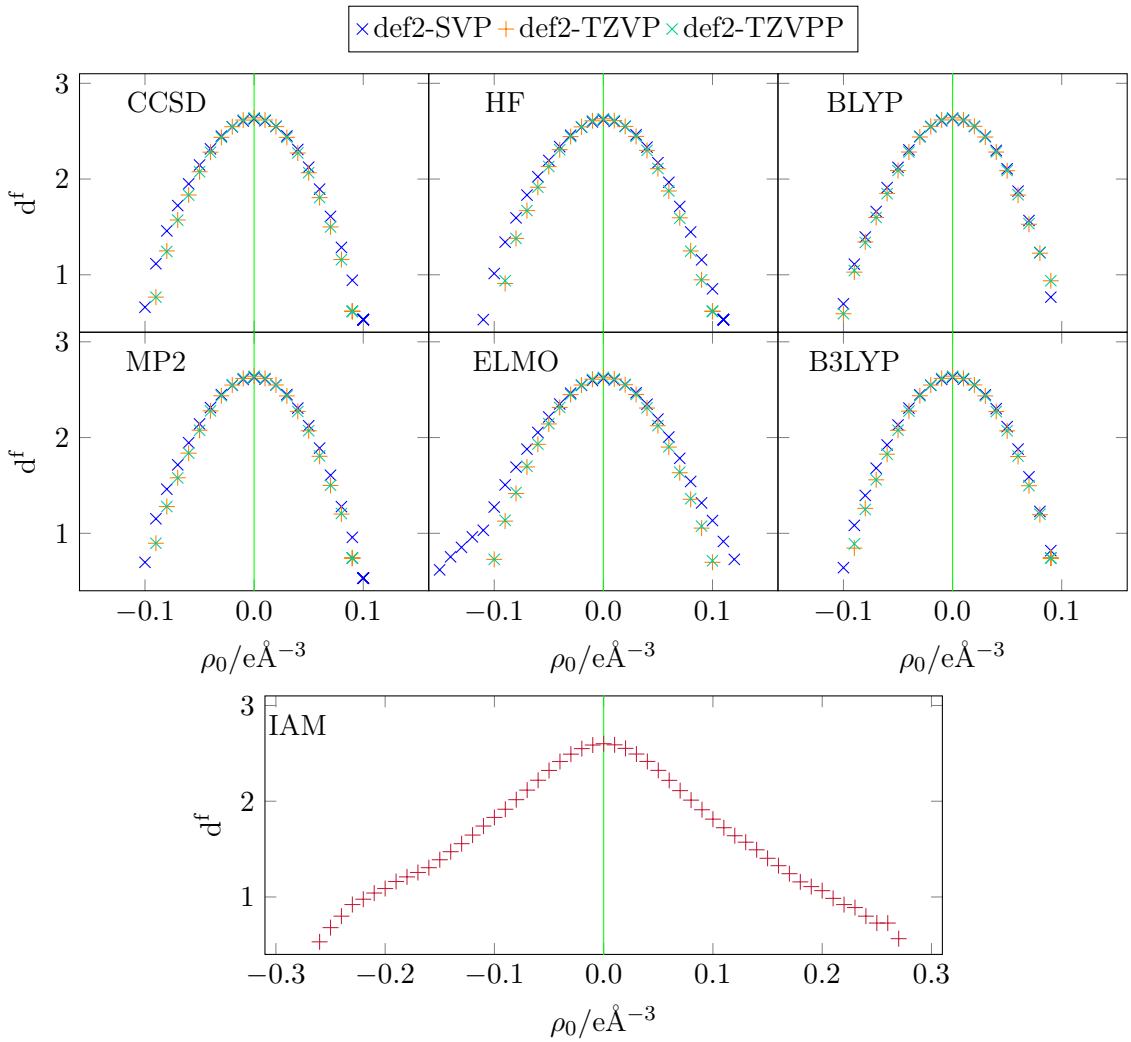


Figure 4.2: Fractal dimension distribution corresponding to the Hirshfeld atom and independent atom model refinements performed on *L*-alanine. Adapted with permission from reference [221]. Copyright 2020 Elsevier B.V.

theoretical with experimental results. Instead, the obtained results are only closer to the "exact results" that can be obtained within the approximations of a particular method.[221]

The main goal of this study was to evaluate the effect of different QM methods, particularly of the post-HF ones on the refinements. Therefore, the resulting differences in the structure factor based quantities will now be discussed in detail. Starting again with analyzing the R and χ^2 values in Table 4.1, it is evident that the trends vary depending on the basis set. For the def2-SVP set of basis functions, the DFT functionals yield the lowest R and χ^2 values, while the results obtained with post-HF methods are higher, and the values resulting from HAR-HF and HAR-ELMO refinements are the highest. In particular, χ^2 approximately takes a value of 1.36 for the DFT functionals, 1.42 for the post-HF methods, 1.63 for HAR-HF and 1.87 for HAR-ELMO. A different trend is observed for the two triple-zeta basis sets, where the best R and χ^2 values are obtained for HAR-CCSD, HAR-B3LYP and HAR-MP2 refinements, intermediate ones for HAR-BLYP and HAR-HF and the highest ones for HAR-ELMO refinements.

The maps of the residual density in the plane of the carboxylate group of *L*-alanine are similar for the post-HF and DFT methods, while for HAR-HF and HAR-ELMO systematically larger regions of negative residual densities (red contour lines) are observed for all three basis sets. This is also confirmed by the Meindl-Henn plots, which are depicted again in Figure 4.3 for a better comparison of the different methods. From this figure, it can be observed that the HAR-HF and HAR-ELMO refinements yield broader Meindl-Henn plots than the other refinement strategies. This is especially the case for the def2-SVP basis set. Moreover, the plot resulting from the HAR-ELMO refinement has a shoulder in the negative region for this basis set that vanishes for the triple-zeta sets of basis functions, leading to very similar plots for the HAR-HF and HAR-ELMO strategies. Except for the HAR-ELMO/def2-SVP plot, all plots follow a parabolic distribution.

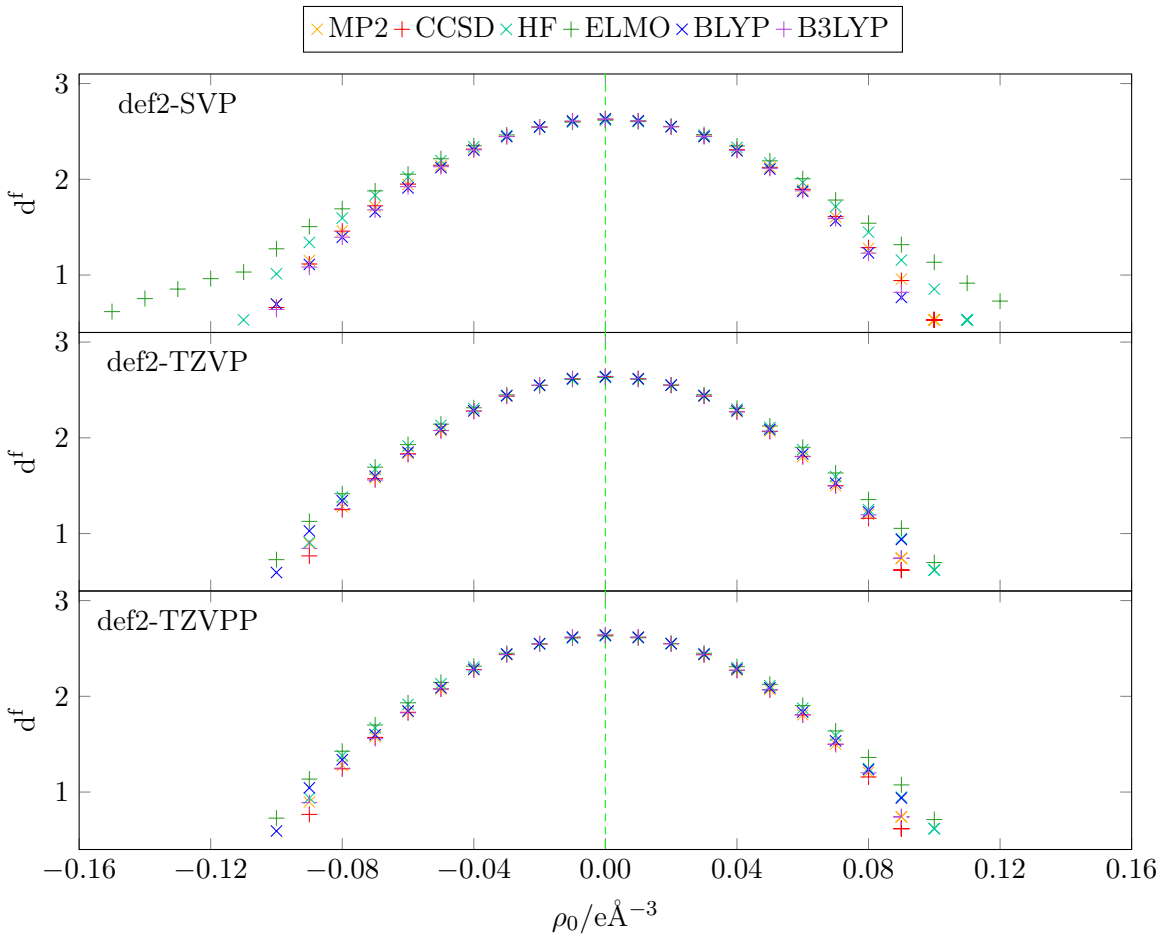


Figure 4.3: Fractal dimension distribution corresponding to the HARs performed on *L*-alanine.

In addition to the residual densities, it is common practice to also compare the so-called deformation densities. These densities are computed by subtracting the promolecular density (that is used in the IAM) from the density obtained with one of the more advanced refinement strategies described in Section 2.3. For the refinements in this study, they are shown in Figure A.1. The positive deformation density corresponds to bond and lone pair regions, or in other words to those areas where additional density is present when changing from the IAM to the aspherical model. In contrast, the negative deformation densities correspond mainly to

the core regions, from where the density is taken away in the aspherical models. Therefore, the deformation densities are often used to judge if the densities resulting for example from the multipole model are chemically meaningful. However, for HARs that are always based on QM calculations this must be the case, and it is indeed so, as it can be seen in Figure A.1.

In summary, if both the effects of the QM methods and the basis sets are considered simultaneously, the highest agreement between calculated and observed structure factors is obtained for HAR-CCSD with the two triple-zeta basis sets, although for these two sets of basis functions HAR-B3LYP and HAR-MP2 refinements yield similar figures of merit. All in all, also considering the results for the residual densities, the best agreement between measured and computed structure factors is obtained for those HARs that are based on post-HF methods and triple-zeta basis sets.

4.3.2 Bond lengths involving hydrogen atoms

As outlined in Section 2.3.4, the main advantage of performing HARs is that the element-hydrogen bond lengths are usually in much better agreement with the neutron reference values than the ones obtained from IAM refinements. This is often shown for average bond lengths.[149,228,229] However, in *L*-alanine only four C−H and three N−H bond lengths are present. Therefore, we decided to compare the individual HAR bond lengths directly with the corresponding neutron ones. This is shown in Figure 4.4, where the values would lie on the diagonal of each plot for a perfect agreement with the neutron bond lengths. The adopted labeling scheme for each atom in *L*-alanine is given in Figure 4.5. For comparison with other studies, in Table A.1 and Table A.2 the average values, average mean absolute differences and average ratios between X-ray and neutron bond lengths are also reported for C−H and N−H bonds, respectively.

In general, from Figure 4.4, it can be noted that the experimental errors associated with all the HAR E−H bond lengths are significantly larger than the ones obtained from refinements of neutron data. Comparing the C−H with the N−H bond lengths depicted on the left and right side of Figure 4.4, respectively, the C−H bond lengths are generally closer to the diagonal than the N−H bond lengths. In fact, the lengths of the N−H bonds are underestimated by all HARs performed in this study. This can probably be explained by the fact that all these bonds are involved in hydrogen bonds with the surrounding molecules in the crystal structure. Since the underlying QM calculations are performed on single molecules only, these hydrogen bonds are not considered in the calculations. As mentioned in Section 2.3.4 and in the previous chapter, the agreement between the E−H bond lengths can be improved by using a cluster of point charges and dipoles in the wavefunction calculations.[227,228,261,298] Another possibility for improving the E−H structures is to use a QM embedding method for the computations of the wavefunction. An example of such an embedded HAR based on the QM/ELMO technique is presented in reference [329] and in Chapter 5. Nevertheless, it is worth stressing again that the HAR bond lengths are significantly closer to the neutron reference values than the IAM bond lengths, which are listed in the caption of Figure 4.4.

For the following discussions about the effects due to the choice of basis set and QM method it is important to keep in mind that conclusions cannot only be drawn from the absolute values of the bond lengths but that also their associated experimental errors need to

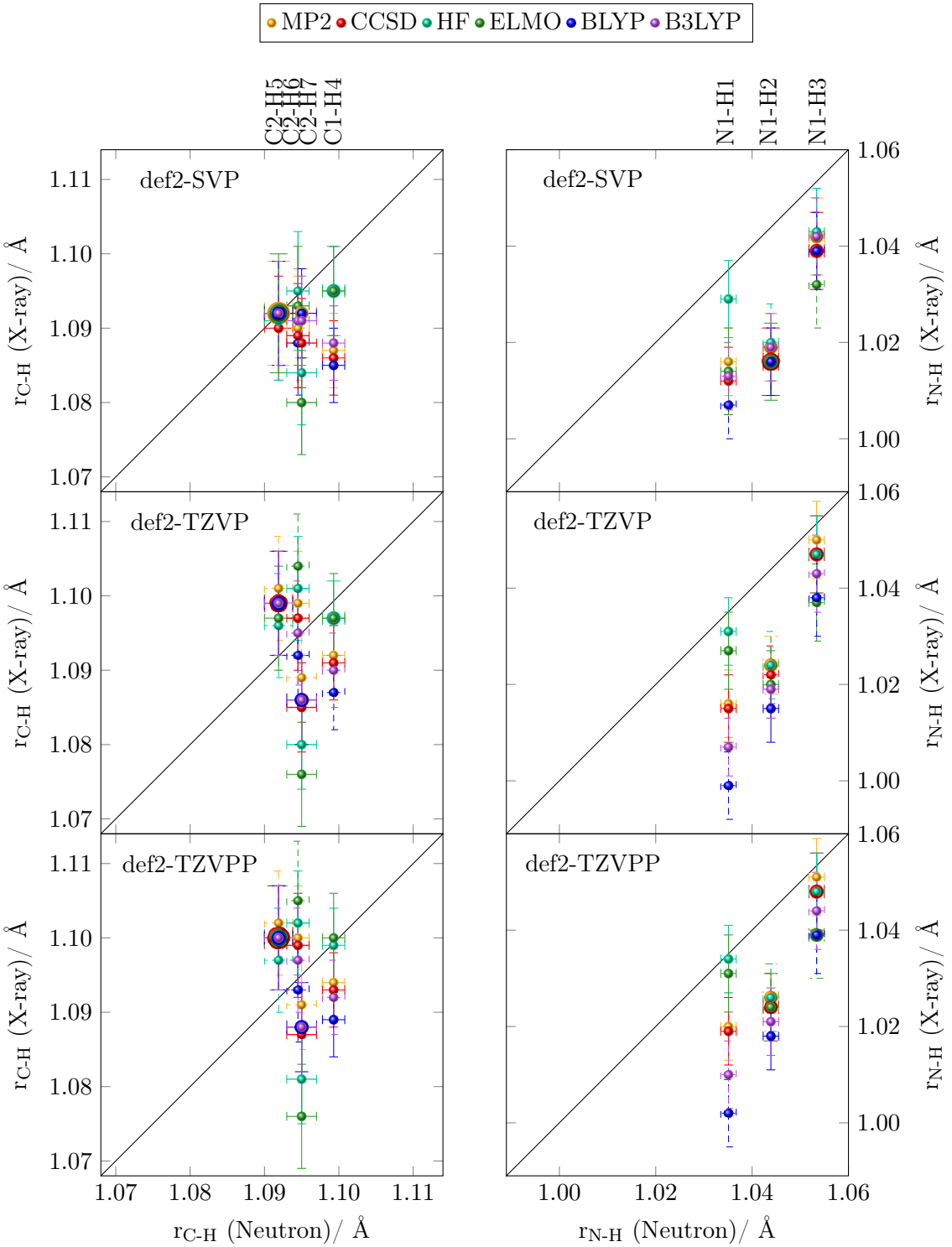


Figure 4.4: E−H bond lengths in *L*-alanine for HARs using the def2-SVP (top), def2-TZVP (middle) and def2-TZVPP (bottom) basis sets plotted against neutron E−H bond lengths with C−H bonds on the left and N−H bonds on the right side. Vertical bars indicate the experimental errors for the X-ray bond lengths, while the horizontal ones represent the uncertainties for the neutron bond lengths. Corresponding E−H bond lengths from the IAM refinement (in Å): N1-H1: 0.935(11), N1-H2: 0.917(10), N1-H3: 1.042(13), C1-H4: 0.999(9), C2-H5: 0.973(12), C2-H6: 0.997(11), C2-H7: 1.021(11). For labels of atoms see Figure 4.5. Adapted with permission from reference [221]. Copyright 2020 Elsevier B.V.

be taken into account. In this study, almost all the E−H bond lengths overlap within their uncertainties and all the presented trends should be also seen in light of this aspect.

For the analysis of the effect of the different basis sets, the C−H and N−H bond lengths will be discussed separately. For the former, it can be observed that three C−H bond lengths are shorter for refinements using def2-SVP than for those with def2-TZVP, while the C2−H7 bond is longer for refinements with the double-zeta than with the triple-zeta basis set. For the N−H bond lengths, no general trend can be observed, except that the bond lengths obtained with different methods are closer to each other for def2-SVP than for the triple-zeta basis sets. The E−H bond lengths obtained with the two triple-zeta sets of basis functions follow very similar trends, although def2-TZVPP yields longer bond lengths.

Concerning the dependence of the E−H bond lengths on the different QM methods, we can notice the following trends. For the most polarized bonds (the three N−H and the two C−H bonds that are in short contacts with oxygen atoms, namely C1−H4 and C2−H6), HAR-HF systematically yields the longest bond lengths that are in most cases also the closest to the neutron reference values. Moreover, for these bonds and only analyzing the triple-zeta basis sets, the DFT functionals yield the shortest bond lengths, while the post-HF methods provide bond lengths that lie in between the values obtained with DFT and those resulting from HAR-HF and HAR-ELMO. Independently of the chosen basis set, the C−H bond lengths resulting from HAR-ELMO refinements are usually the closest to the HF ones, while the corresponding N−H bond lengths differ more strongly.

All in all, it should be stressed again that although some trends can be identified for the E−H bond lengths, almost all of them overlap within their standard deviations. Therefore, due to these uncertainties, the main conclusion that can be drawn from the comparison of the individual bond lengths in Figure 4.4 is that the differences between the adopted QM methods are not statistically significant.

4.3.3 Atomic displacement parameters

In Section 2.1.10, it was already mentioned that obtaining ADPs for hydrogen atoms is desirable but difficult. Here, the ADPs obtained for all the different HARs will be compared to the neutron reference values. The ADPs can be visually compared in Figure 4.5, where all the refined crystal structures of *L*-alanine are depicted for the performed HARs. Additionally, the IAM and neutron reference structures are also shown.

Since six individual parameters are independently refined for each atom, the comparison of average values is more meaningful than for the previously described bond lengths. Therefore, in this section, the average mean absolute differences and average ratios between X-ray and neutron ADPs will be discussed separately for hydrogen atoms bonded to carbon and nitrogen, for which the results obtained from the statistical analysis are given in Tables 4.2 and 4.3, respectively. The same quantities are also reported for the non-hydrogen ADPs in Table A.3.

From the visual inspection of Figure 4.5, it can be observed that the hydrogen ADPs resulting from HAR are longer and flatter than the ones in the neutron structure. For different methods and basis sets the general shape of the hydrogen ADPs remains, although their sizes and orientations vary. Similar to the previously described trends for the structure factor based quantities and the E−H bond lengths, the discrepancies for hydrogen ADPs are larger

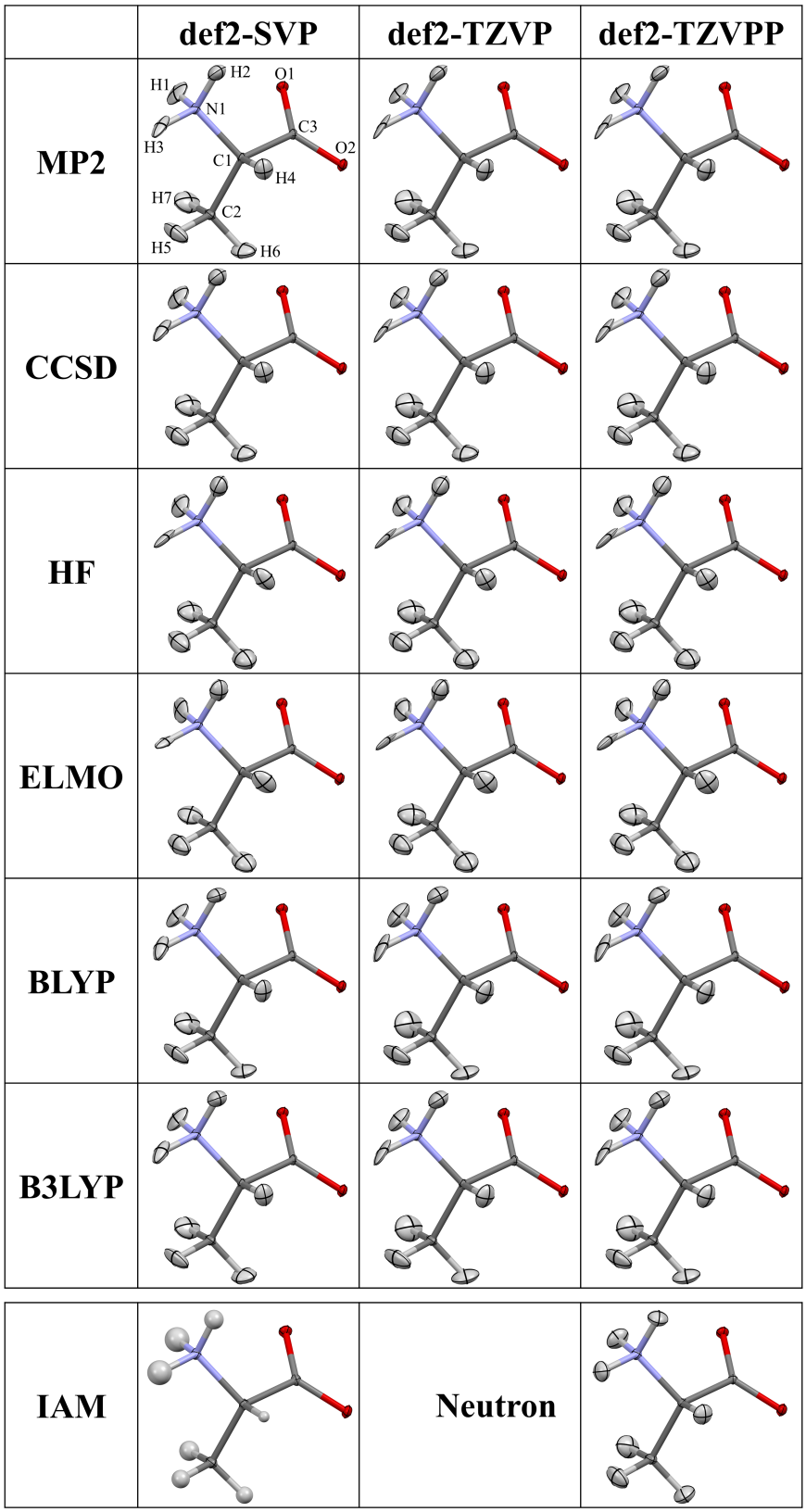


Figure 4.5: Refined crystal structures of *L*-alanine obtained from the refinement of neutron data[149] and from HAR and IAM refinements of X-ray data.[244] All atomic displacement parameters are depicted at the 50% probability level. Adapted with permission from reference [221]. Copyright 2020 Elsevier B.V.

Table 4.2: Statistical analysis of the HAR ADPs for hydrogen atoms bonded to carbon: (i) mean ratios of the diagonal elements $\langle U^{ii}_{\mathrm{HAR}}/U^{ii}_{\mathrm{neutron}}\rangle$, (ii) mean absolute differences of the diagonal terms $\langle|U^{ii}_{\mathrm{HAR}} - U^{ii}_{\mathrm{neutron}}|\rangle$ and (iii) mean absolute differences of the non-diagonal elements $\langle|U^{ij}_{\mathrm{HAR}} - U^{ij}_{\mathrm{neutron}}|\rangle$. For each basis set, the first column refers to the value of the quantity and the second column to the corresponding standard deviation upon averaging. Adapted with permission from [221]. Copyright 2020 Elsevier B.V.

$\langle U^{ii}_{\mathrm{HAR}}/U^{ii}_{\mathrm{neutron}}\rangle$/Å						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	1.02	0.17	1.07	0.24	1.10	0.24
CCSD	1.09	0.16	1.14	0.21	1.16	0.22
HF	1.21	0.21	1.28	0.18	1.29	0.19
ELMO	1.20	0.30	1.31	0.22	1.32	0.23
BLYP	1.02	0.20	1.17	0.31	1.18	0.31
B3LYP	1.03	0.16	1.16	0.26	1.17	0.26

$\langle\|U^{ii}_{\mathrm{HAR}} - U^{ii}_{\mathrm{neutron}}\|\rangle$/Å^2						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.0035	0.0019	0.0049	0.0045	0.0051	0.0048
CCSD	0.0033	0.0026	0.0049	0.0048	0.0053	0.0050
HF	0.0043	0.0032	0.0064	0.0038	0.0066	0.0042
ELMO	0.0054	0.0043	0.0068	0.0051	0.0071	0.0054
BLYP	0.0038	0.0030	0.0068	0.0066	0.0069	0.0067
B3LYP	0.0032	0.0023	0.0059	0.0059	0.0060	0.0061

$\langle\|U^{ij}_{\mathrm{HAR}} - U^{ij}_{\mathrm{neutron}}\|\rangle$/Å^2 with $i \neq j$						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.0033	0.0026	0.0029	0.0025	0.0031	0.0025
CCSD	0.0033	0.0027	0.0028	0.0025	0.0029	0.0025
HF	0.0049	0.0031	0.0040	0.0024	0.0041	0.0022
ELMO	0.0063	0.0036	0.0051	0.0027	0.0052	0.0023
BLYP	0.0032	0.0026	0.0039	0.0025	0.0040	0.0025
B3LYP	0.0031	0.0024	0.0032	0.0025	0.0033	0.0026

between the double-zeta and the triple-zeta basis sets, while the hydrogen ADPs obtained with the two triple-zeta sets of basis functions are very similar to each other.

In addition to the visual comparison, a statistical analysis has been performed for the hydrogen ADPs (compare Tables 4.2 and 4.3). The ratios between the diagonal-elements of the HAR and neutron ADPs are in all cases above 1.00, meaning that the HAR ADPs are on average larger than the neutron ones. Furthermore, the mean absolute differences between

Table 4.3: Statistical analysis of the HAR ADPs for hydrogen atoms bonded to nitrogen: (i) mean ratios of the diagonal elements $\langle U^{ii}_{\mathrm{HAR}}/U^{ii}_{\mathrm{neutron}}\rangle$, (ii) mean absolute differences of the diagonal terms $\langle|U^{ii}_{\mathrm{HAR}} - U^{ii}_{\mathrm{neutron}}|\rangle$ and (iii) mean absolute differences of the non-diagonal elements $\langle|U^{ij}_{\mathrm{HAR}} - U^{ij}_{\mathrm{neutron}}|\rangle$. For each basis set, the first column refers to the value of the quantity and the second column to the corresponding standard deviation upon averaging. Adapted with permission from [221]. Copyright 2020 Elsevier B.V.

$\langle U^{ii}_{\mathrm{HAR}}/U^{ii}_{\mathrm{neutron}}\rangle$/Å						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	1.23	0.68	1.24	0.59	1.30	0.61
CCSD	1.32	0.70	1.33	0.63	1.39	0.65
HF	1.45	0.97	1.48	0.88	1.52	0.88
ELMO	1.40	1.16	1.46	1.01	1.50	0.98
BLYP	1.21	0.46	1.37	0.42	1.41	0.43
B3LYP	1.22	0.55	1.34	0.49	1.38	0.49

$\langle\lvert U^{ii}_{\mathrm{HAR}} - U^{ii}_{\mathrm{neutron}}\rvert\rangle$/Å^2						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.0075	0.0065	0.0068	0.0055	0.0071	0.0057
CCSD	0.0080	0.0064	0.0073	0.0057	0.0077	0.0061
HF	0.0109	0.0086	0.0093	0.0083	0.0094	0.0086
ELMO	0.0143	0.0095	0.0110	0.0092	0.0106	0.0093
BLYP	0.0064	0.0036	0.0065	0.0057	0.0072	0.0058
B3LYP	0.0067	0.0049	0.0064	0.0051	0.0067	0.0054

$\langle\lvert U^{ij}_{\mathrm{HAR}} - U^{ij}_{\mathrm{neutron}}\rvert\rangle$/Å^2 with $i \neq j$						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.0074	0.0065	0.0085	0.0079	0.0086	0.0081
CCSD	0.0071	0.0063	0.0081	0.0075	0.0081	0.0077
HF	0.0076	0.0068	0.0079	0.0078	0.0079	0.0077
ELMO	0.0079	0.0058	0.0078	0.0068	0.0076	0.0065
BLYP	0.0069	0.0057	0.0082	0.0084	0.0083	0.0084
B3LYP	0.0063	0.0060	0.0078	0.0079	0.0077	0.0080

neutron and HAR ADPs are not zero, which remains the case for most parameters also when considering the corresponding standard deviations. Moreover, comparing the results for the hydrogen atoms bonded to carbon with those bonded to nitrogen, it can be seen that the former are generally in better agreement with the neutron values than the latter. In the original publication of this study (reference [221]) we proposed that the description of the ADPs could be improved by using cluster charges or the ELMO embedded HAR strategy.[329]

This aspect will be further discussed in Chapter 5.

All the trends described in the paragraph above are general and not affected by the choice of QM method or basis set. Nevertheless, some trends can be observed regarding their individual influences on the ADPs. Let us first focus on the effect of the basis set, where trends can only be observed for the diagonal ADPs of hydrogen atoms bonded to carbon and non-diagonal ADPs of hydrogen atoms bonded to nitrogen. In these cases, the discrepancies with neutron ADPs are smaller for def2-SVP than for the triple-zeta basis sets. Moreover, for both the diagonal and non-diagonal ADPs of hydrogen atoms bonded to carbon and the non-diagonal ADPs of hydrogen atoms bonded to nitrogen, the results obtained for the triple-zeta basis sets are generally very similar.

Furthermore, in Tables 4.2 and 4.3 some trends for the effect of the QM method can be also observed. When def2-SVP is used, the refinements based on post-HF and DFT methods provide ADPs characterized by a better agreement with the neutron reference values than HAR-HF or HAR-ELMO. However, a clear ranking between the post-HF and DFT methods is difficult to establish. Only for hydrogen atoms bonded to nitrogen, it is possible to observe that the two DFT functionals give lower discrepancies than the two post-HF methods, especially compared to CCSD. For the two triple-zeta basis sets, different trends are obtained for the hydrogen atoms bonded to carbon or nitrogen. For all the hydrogen atoms bonded to carbon, the worst agreement between HAR and neutron ADPs is obtained for the HAR-ELMO refinements. Intermediate results are obtained for HARs based on the HF or BLYP methods, while the post-HF techniques lead to the lowest discrepancies. Pertaining to the ADPs of hydrogen atoms bonded to nitrogen, for the diagonal elements, HAR-ELMO and HAR-HF give the worst results, while the application of post-HF methods and DFT functionals improves the agreement. However, establishing a clear trend between the correlated methods is again difficult.

All in all, it seems that HAR-HF and HAR-ELMO refinements lead to displacement parameters for hydrogen ADPs that are the most different compared to the neutron ones than the other methods. However, all the discussed differences between the HARs are very small and statistically not significant due to the large standard deviations. Moreover, all the hydrogen ADPs obtained with HAR show differences to the neutron ADPs irrespective of the chosen methods or basis sets.

As mentioned above, a corresponding statistical analysis has also been performed for the non-hydrogen ADPs, for which the results are collected in Table A.3. For the diagonal elements of the non-hydrogen ADPs, HARs based on HF calculations yield the best agreement with the neutron values, regardless of the chosen basis set. Furthermore, the post-HF methods are better than the DFT functionals. For HAR-ELMO no clear trend can be established. Concerning the non-diagonal ADPs, the best agreement is obtained for HAR-CCSD refinements, while the worst results correspond to the HAR-ELMO refinements.

4.3.4 Electron densities

With the goal of explaining the results obtained from the different HARs, in this subsections we compare differences in the electron densities calculated with all the QM methods and basis sets used in this study. In particular, we base our comparison on electron densities obtained

for the common starting point of all the HARs in this study, namely the IAM geometry. This structure was chosen because it corresponds to the geometry used in the first HAR iteration that is the most crucial one in the iterative HAR procedure. The reason for this choice is that the first iteration directs the refinement towards the final converged structure and the aspherical electron density obtained at this first step has the largest influence on the outcome of the refinement. For instance, this can be seen from the very first HAR study, where the refinement was not based on an iterative procedure, but only one HAR iteration was performed.[227] Despite this approximation, promising structural parameters were already obtained.[227] Therefore, in the work presented in this chapter, we computed electron density differences for the IAM structure subtracting a reference density from all the other densities. In particular, for the comparison of the different QM methods, the HF density was used as reference, while for estimating the effect of the basis set, the densities obtained with the def2-SVP basis set were the benchmarks. The resulting difference densities for the basis set comparison are shown in Figure 4.6, while those for the comparison of the different QM methods are reported in Figure 4.7.

The difference densities in Figure 4.6 show that the choice of the basis set has an influence on the electron density of the complete molecule, even in the areas where the hydrogen atoms are located. It is worth noting that the difference densities obtained with the two triple-zeta basis sets are very similar, which explains the previously observed similarities between the HARs exploiting the def2-TZVP and def2-TZVPP basis sets.

Contrary to the previously discussed effect of the basis set, the choice of the QM method mainly affects the electron densities of the non-hydrogen atoms (compare Figure 4.7). While the MP2 and CCSD electron densities differ from the HF ones almost only at the carboxylate group, differences for the DFT densities are observed for all non-hydrogen atoms. This is consistent with the previously described trends for the diagonal elements of the non-hydrogen ADPs, for which similarities between the post-HF HARs (MP2 and CCSD) and between the DFT ones (B3LYP and BLYP) were observed.

Furthermore, from Figure 4.7 it appears that the largest differences are obtained for the ELMO densities. However, it is important to note that the ELMOs were not tailor-made for the IAM geometry, but were transferred to it from a model molecule with the neutron geometry. Therefore, the ELMOs localized on the E−H bonds are not optimized for the too short bond lengths in the IAM structure and require a larger renormalization. This leads to the significant deformations in the densities. In fact, if the difference density is computed for the neutron structure as shown in Figure 4.8, significantly smaller differences are obtained. Compared to the IAM structure, the negative difference densities located on the hydrogen atoms and some of the E−H bonds vanish for the neutron structure and only some differences remain for the two oxygen atoms.

In summary, the fact that the differences in the electron densities obtained for the refinements based on the different QM methods are not located on the hydrogen atoms explain why the corresponding structural parameters are very similar for the different QM methods and why it was sometimes difficult to find clear trends.

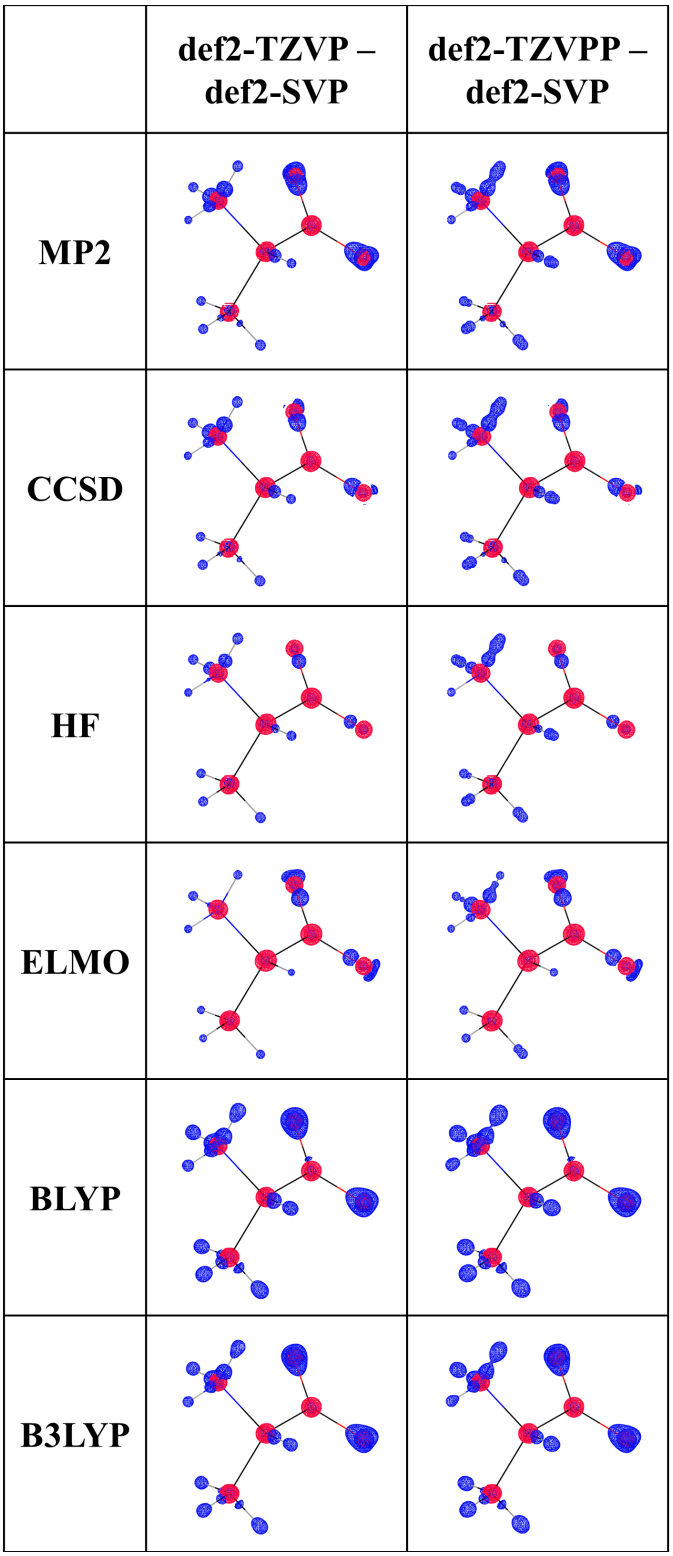


Figure 4.6: Comparison of the electron densities obtained with the different basis sets (references: def2-SVP electron distributions) for each exploited QM method. All electron densities were computed on the IAM structure. Contour level: 0.02 e bohr^{-3}; colors in the difference density plots: blue (positive) and red (negative); the orientation of the molecules corresponds to the one in Figure 4.5. Adapted with permission from reference [221]. Copyright 2020 Elsevier B.V.

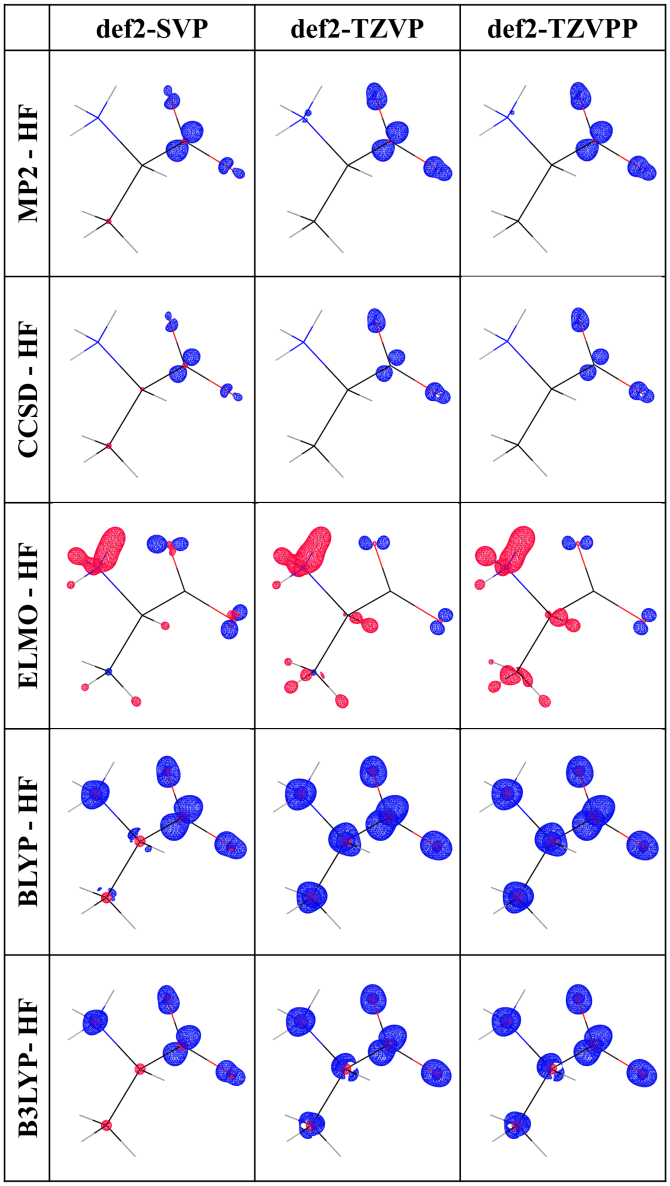


Figure 4.7: Comparison of the electron densities obtained with the different QM methods (references: HF electron distributions) for each of the used basis sets. All electron densities were computed on the IAM structure. Contour level: 0.02 e bohr^{-3}; colors in the difference density plots: blue (positive) and red (negative); the orientation of the molecules corresponds to the one in Figure 4.5. Adapted with permission from reference [221]. Copyright 2020 Elsevier B.V.

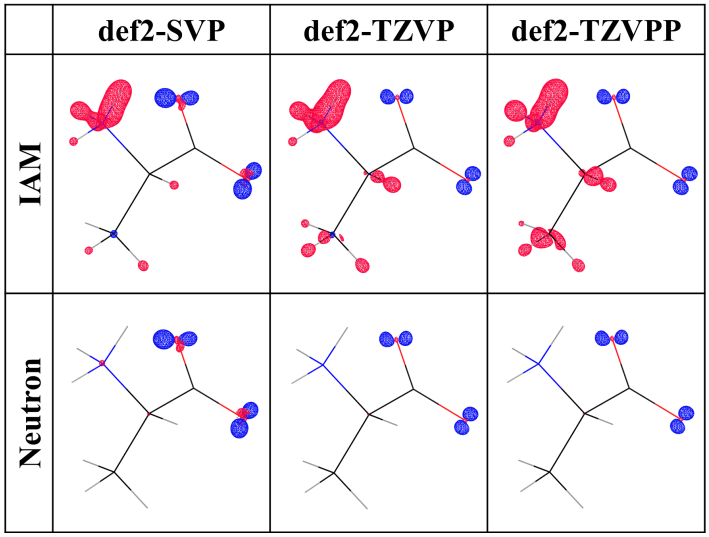


Figure 4.8: Electron density differences at ELMO level with the HF electron distributions as references for each adopted basis set and computed on the IAM or neutron structures. Contour level: 0.02 e bohr^{-3}; colors in the difference density plots: blue (positive) and red (negative); the orientation of the molecules corresponds to the one in Figure 4.5.

4.4 Conclusions: are post-HF methods necessary for HAR?

In this chapter, the first Hirshfeld atom refinements based on the post-HF techniques MP2 and CCSD were presented and compared to HARs exploiting ELMO, HF and DFT methods. All refinements were performed on the crystal structure of *L*-alanine.

Evaluating the necessity for post-HF methods for HARs on the *L*-alanine structure, we found that post-HF techniques do not systematically improve the structural parameters in comparison to other QM methods. All in all, some differences could be observed for the structural parameters (E−H bond lengths and ADPs) resulting from HARs based on theoretical calculations with different levels of theory or basis sets. However, although from these differences some individual trends could be derived, it was difficult to establish more general rankings between the QM methods because the results seem to depend on several aspects, for example the crystal environment of the hydrogen atoms. Moreover, in most cases, the differences are too small to draw clear conclusions and are much smaller than the standard deviations corresponding to the structural parameters. Therefore, from the results obtained in this study, it could be concluded that post-HF methods may not be necessary for the determination of bond lengths involving hydrogen atoms or hydrogen ADPs. However, before definitive conclusions can be drawn, further investigations on other systems may be required, for example on compounds containing heavy elements. Furthermore, the performances of post-HF methods for HARs could be evaluated again when the crystal environment is taken explicitly into account. In fact, this was done in a later study by Chodkiewicz *et al.*[330] For the structures of urea and oxalic acid, they performed HARs based on the five QM methods that were used in the present study (except ELMOs) exploiting the basis set cc-pVTZ and a 8 Å cluster of multipoles. They concluded that refinements based on B3LYP, MP2 or CCSD provide the most promising results in comparison to the traditional HAR-HF or HAR-BLYP

refinements.[330]

Concerning the figures of merit that are a measure for the agreement between calculated and observed structure factors, the following ranking of the different levels of theories could be established (from the worst to the best agreement): ELMO, HF, BLYP, MP2 $\approxeq$ B3LYP, CCSD. This ranking is consistent with the general performances of these QM methods and indicates that the figures of merit could be used to test the quality of different QM methods. Therefore, the current study could be seen as a starting point, and the presented comparison of different types of HARs could be continued based on a database of high-quality diffraction datasets. In fact, such a database would not only be highly useful for the validation of new refinement strategies but also for benchmarking different QM methods by comparing the experimental and calculated structure factors. In this sense, such data could be valuable beyond the purpose of structural refinements.

5 An ELMO embedding strategy for more accurate E–H bond lengths

5.1 Introduction

In crystal structures, each molecule is surrounded by a number of symmetry generated copies of itself. If this crystal environment is considered in the QM calculations underlying the Hirshfeld atom refinements, the obtained E−H bond lengths become closer to the neutron reference values.[227,228,298] In the software *TONTO*, this is achieved by generating a cluster of molecules surrounding the QM region. In particular, *TONTO* considers only those molecules with at least one atom lying within a specified radius (default is 8 Å) around the central QM unit. After the cluster is generated, charges and dipoles are computed for the Hirshfeld atoms obtained from a previous QM calculation (compare HAR procedure, Section 2.3.4). The resulting charges are placed directly at the positions of the surrounding atoms in the cluster. To also simulate the dipoles associated with the Hirshfeld atoms, two additional charges with opposite signs are placed at a distance of ±0.001 a.u. on each side of the different atoms.[227]

Despite the procedure described above can improve the E−H bonds lengths, different studies[228,298] reported that the lengths of polar E−H bonds, which are involved in short intermolecular contacts, are still systematically shorter than the neutron reference values. Therefore, Capelli *et al.* proposed to also include the nearest surrounding molecules in the QM calculations.[228] This approach has recently been tested by Chodkiewicz *et al.*[330] In particular, they used two sizes for the surrounding cluster, where in addition to the molecular reference unit also the following atoms were included in the wavefunction calculations: (i) those of the molecules that directly form hydrogen bonds with the molecular reference unit; or (ii) those belonging to all the molecules that lie within a 3.5 Å radius around the reference unit. Additionally, both clusters were surrounded by point charges within an 8 Å radius. For both cluster sizes, Chodkiewicz *et al.* performed Hirshfeld atom refinements of urea and oxalic acid. They reported that the computational cost increased significantly. However, they could not observe a better agreement with the neutron references for the resulting E−H bond lengths compared to those obtained from HARs with a cluster of charges and dipoles. Nevertheless, they also pointed out that neither of the studied structures contains very strong interactions with the molecules surrounding the corresponding reference units. Therefore, they could not definitely rule out the usefulness of an explicit quantum mechanical treatment of the surrounding molecules for systems characterized by very strong intermolecular interactions.[330]

To further investigate this possibility, we implemented an alternative approach that consists in embedding the central QM unit into an environment of frozen ELMOs. To this purpose, we coupled the QM/ELMO and QM/ELMO/MM techniques (see Section 1.5) with HAR. The corresponding ELMO-embedded-HAR strategy and most of the results shown in this chapter have been previously published in reference [329].

5.2 Computational details

The following five types of HAR were performed:

(i) a traditional HAR without embedding, hereinafter labeled as "no embedding";

(ii) HARs using two different clusters of point charges with radii of 4 and 8 Å, respectively labeled as "4 Å charges" and "8 Å charges";

(iii) HARs using two different clusters of point charges and dipoles with radii of 4 and 8 Å, respectively labeled as "4 Å charges & dipoles" and "8 Å charges & dipoles";

(iv) HARs based on QM/ELMO calculations, where the QM part corresponds to the molecular reference unit and ELMOs are transferred to all the molecules with at least one atom within a 4 or 8 Å radius (compare Figure 5.1A), respectively labeled as "4 Å ELMOs" and "8 Å ELMOs";

(v) HAR based on QM/ELMO/MM calculations, where the QM part corresponds to the molecular reference unit, the ELMO region extends up to a radius of 4 Å and the MM part includes the molecules between the 4 and 8 Å radii (compare Figure 5.1B), labeled as "4 Å ELMOs + MM (4 Å - 8 Å)".

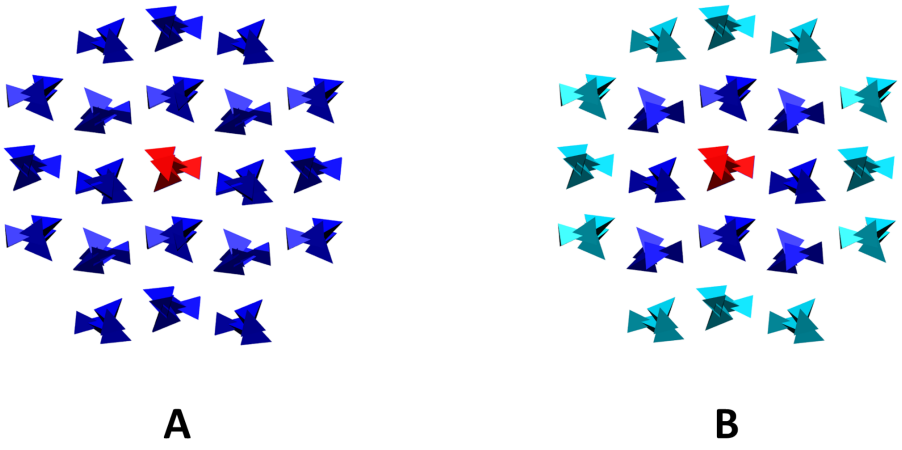


Figure 5.1: Schematic representations (with xylitol molecules in polyhedral rendering) of the crystal environment treatment in the ELMO-embedded Hirshfeld atom refinements: (A) QM/ELMO and (B) QM/ELMO/MM cases. Red: central QM part; blue: surrounding ELMO part; cyan: optional surrounding MM part. Reprinted with permission from reference [329]. Copyright 2021 American Chemical Society.

Note that the difference between type (ii) and type (iii) is that in the former only the Hirshfeld charges are used, while in the latter charges that mimic dipoles are also added. For the refinements of type (iv) and (v) the simple point charge models are replaced by a fully or partial QM treatment of the surrounding molecules. To this purpose, the QM/ELMO and QM/ELMO/MM embedding strategies (compare Section 1.5) are exploited. Either a QM/ELMO or a QM/ELMO/MM calculation is performed, where the molecular reference unit is still treated at QM level, but the symmetry-generated molecules of the surrounding cluster are described using either only ELMOs or ELMOs in combination with an outer layer of MM charges (compare Figure 5.1). In Table 5.1 the number of atoms for each part (QM, ELMO and, if available, MM part) and basis functions for each of the different cluster sizes

Table 5.1: Number of atoms in the different regions and overall number of basis functions used in the different ELMO-embedded Hirshfeld atom refinements of xylitol.

	QM/ELMO 4 Å ELMOs	QM/ELMO 8 Å ELMOs	QM/ELMO/MM 4 Å ELMOs + MM (4 Å - 8 Å)
Number of atoms belonging to the different regions			
QM region	22	22	22
ELMO region	352	1232	352
MM region	n/a	n/a	880
Total	374	1254	1254
Number of basis functions used in the refinements			
cc-pVDZ	3570	11970	3570
cc-pVTZ	9010	30210	9010

are reported.

The refinements of type (i) to (iii) were carried out using the interface *lamaGOET*[300] (for details about *lamaGOET*, see previous chapters). For the remaining two types we used an in-house *bash* script that follows the philosophy of *lamaGOET*. Our *bash* script interfaces the program *TONTO*[297] with a modified version of *Gaussian09*,[150] where the QM/ELMO and QM/ELMO/MM strategies have been implemented. The electron density is passed from *Gaussian09* to *TONTO* using a formatted checkpoint file. To process the formatted checkpoint files that include information on the ELMO and ELMO/MM embeddings, it was necessary to slightly modify the *TONTO* code. For the QM/ELMO/MM calculations, our *bash* script further interfaces the software *AMBER 2016*[148] with the modified version of *Gaussian09*. For the optional MM layer, the general AMBER force field (GAFF)[331] with charges from the AM1-BCC model[332,333] was exploited. Irrespective of the type of embedding, all the underlying QM calculations were performed using the B3LYP[306–309] functional in combination with the basis sets cc-pVDZ and cc-pVTZ.[145]

Using the five different types of HAR, the structure[263] of xylitol was refined against X-ray diffraction data that were collected by Madsen and coworkers at a temperature of 122 K and up to a resolution of 0.41 Å. The resulting structural parameters were compared to those of the corresponding neutron structure[334], which was refined by Madsen *et al.* by exploiting neutron data measured at the same temperature of the X-ray data. No cutoff was applied except that only reflections with $F > 0$ were considered in the refinement.

The results for the refinement types (i), (ii), (iv) and (v) have been previously published in reference [329]. However, at the time of that study, the option to also add the dipoles (refinement type (iii)), was not available in *lamaGOET*. The corresponding results have been collected for this thesis.

5.3 Results and discussion

To evaluate whether the ELMO embedding improves the outcome of the Hirshfeld atom refinements, the following quantities were evaluated: (i) figures of merit, values of the residual density in the unit cell, and Meindl-Henn plots;[249] (ii) the lengths of the O−H and C−H bonds in xylitol as well as distances and angles of the intermolecular hydrogen bonds; (iii) the ADPs; and (iv) the electron densities obtained from the different QM calculations.

5.3.1 Structure factor based descriptors

Beginning with the structure factor based descriptors, the R and χ^2 values resulting from the different HARs are reported in Table 5.2. In the same table, the minimum, maximum and root mean square values of the residual density in the unit cell are also given. For all the different

Table 5.2: Figures of merit, and maximum, minimum and root mean square values of the residual density in the unit cell resulting from the different HARs of the xylitol crystal structure.

Basis set and type of embedding	$R(F)$/ %	$R_w(F)$/ %	χ^2	ρ_{max}/ eÅ^{-3}	ρ_{min}/ eÅ^{-3}	ρ_{rms}/ eÅ^{-3}
cc-pVDZ						
no embedding	1.79	1.29	0.67	0.131	-0.129	0.030
4 Å charges	1.78	1.28	0.66	0.130	-0.129	0.030
8 Å charges	1.78	1.28	0.66	0.130	-0.129	0.030
4 Å charges & dipoles	1.78	1.28	0.66	0.128	-0.131	0.030
8 Å charges & dipoles	1.78	1.28	0.66	0.128	-0.131	0.030
4 Å ELMOs	1.80	1.30	0.68	0.128	-0.129	0.031
8 Å ELMOs	1.80	1.30	0.68	0.128	-0.129	0.031
4 Å ELMOs + MM (4 Å - 8 Å)	1.80	1.30	0.68	0.128	-0.129	0.031
cc-pVTZ						
no embedding	1.71	1.21	0.59	0.113	-0.124	0.028
4 Å charges	1.70	1.20	0.58	0.112	-0.123	0.027
8 Å charges	1.70	1.20	0.58	0.112	-0.123	0.027
4 Å charges & dipoles	1.68	1.18	0.56	0.111	-0.122	0.027
8 Å charges & dipoles	1.68	1.18	0.56	0.111	-0.122	0.027
4 Å ELMOs	1.69	1.19	0.57	0.113	-0.119	0.028
8 Å ELMOs	1.69	1.19	0.57	0.112	-0.121	0.028
4 Å ELMOs + MM (4 Å - 8 Å)	1.69	1.19	0.57	0.112	-0.121	0.028

types of refinements, the figures of merit are very low, indicating the very good quality of the X-ray dataset. The values decrease when the cc-pVTZ basis set is used. Within a set of basis functions, the differences between the refinements with and without embedding are very small. Moreover, the different types of embedding always give the same figures of merit. For example, when the basis set cc-pVTZ is used, the following values for χ^2 are obtained: 0.59 for the refinement without embedding, 0.58 for the refinements with charges, 0.56 for HARs with charges and dipoles, and 0.57 for all three refinements with ELMO embedding. Also the minimum and maximum values of the residual density in the unit cell are very low (compare again Table 5.2), and only small improvements can be observed when the crystal

environment is taken into account. This observation is confirmed by the Meindl-Henn plots (see Figure 5.2), which are very narrow and practically equally distributed for all the adopted HAR strategies.

In summary, the figures of merit indicate that for the very good X-ray dataset of xylitol, the traditional HAR is completely sufficient for achieving an excellent agreement between the measured and calculated structure factors. Nevertheless, this does not mean that the structural parameters for the hydrogen atoms cannot be improved by the embedded refinements. In fact, this will be evaluated in the next two subsections.

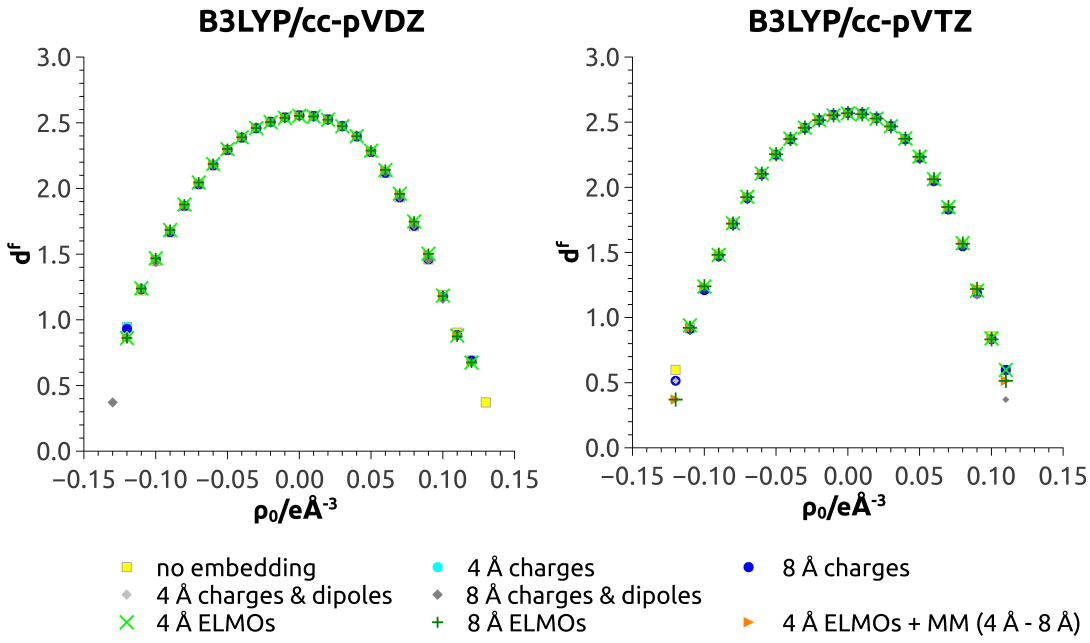


Figure 5.2: Meindl-Henn plots[249] of the fractal distributions associated with the Hirshfeld atom refinements of the xylitol crystal structure.

5.3.2 Bond lengths involving hydrogen atoms

In the crystal structure of xylitol, all the O−H bond lengths are involved in hydrogen bonds, as can be seen in Figure 5.3. Therefore, it can be expected that the traditional HAR and possibly also the HARs with charges and dipoles will yield E−H that are shorter than the neutron reference bonds. However, as mentioned at the beginning of this chapter, the question is whether a fully QM embedding strategy can improve the lengths of bonds that involve hydrogen atoms. To answer this question, the individual E−H bond distances in xylitol are compared (Figure 5.4). Additionally, a statistical analysis of the bond lengths is reported in Table 5.3.

In Figure 5.4, it is interesting to note that the bond lengths are not influenced by the size of the embedding region (4 vs 8 Å) but only by the adopted embedding strategy or the basis set. As expected, the traditional HAR provides the shortest O−H bond lengths, which are significantly shorter than the neutron reference values. A small improvement can be observed for the refinements with cluster charges, which lead to slightly longer O−H bond lengths.

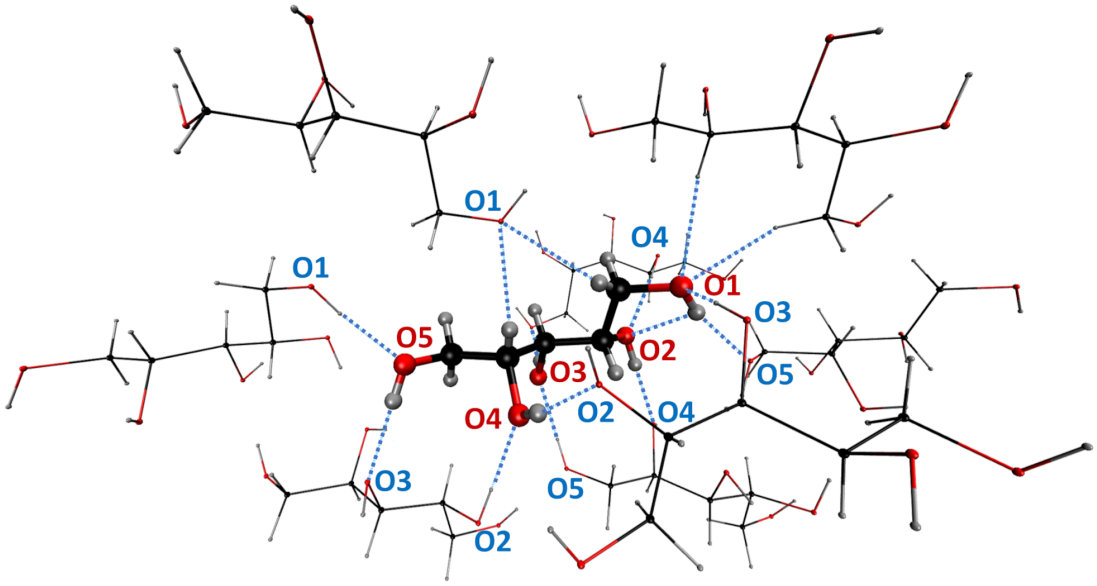


Figure 5.3: Network of inter- and intramolecular contacts in the xylitol crystal structure. The oxygen atoms labeled in red belong to the QM region, while the oxygen atoms labeled in blue are part of the ELMO subsystem.

However, irrespective of the chosen basis set or cluster size, the improvements are statistically not significant. For the embedding with charges and dipoles the O−H bond lengths further elongate, but are nevertheless too short compared to the neutron references. In contrast, these bond lengths improve significantly if the ELMO embedding strategy is adopted. Especially for the basis set cc-pVDZ, an optimal agreement between the ELMO-embedded-HAR and neutron values can be observed for every O−H bond length in the structure of xylitol. For the refinements with the basis set cc-pVTZ, the O3−H13 bond length is overestimated for the ELMO or ELMO/MM embeddings. In this case, the bond lengths resulting from HARs with charges and dipoles are closer to the neutron reference. However, for all the other bond lengths the ELMO-embedded-HAR strategy provides the best agreement with the neutron benchmarks.

The corresponding C−H bond lengths in the xylitol crystal structure are also shown in Figure 5.4. For both basis sets, even eventually observable differences between the different types of HARs are not statistically significant in most cases. In fact, the experimental errors overlap. This observation can be explained by considering the network of intermolecular interactions in the crystal structure of xylitol (compare again Figure 5.3). In fact, most of the C−H bonds are not involved in the intermolecular contacts. Therefore, an embedded QM calculation is not really essential for this type of bonds, which is also consistent with the findings from previous studies[228,298] that generally report a better agreement between the HAR and neutron values for the C−H bond lengths.

The previously described trends can be also observed in the statistical analysis in Table 5.3, where the mean ratios and mean absolute differences between HAR and neutron values for the C−H and O−H bond lengths are reported. Additionally, the same quantities are also given for the hydrogen-acceptor (H···A) distances and the oxygen-hydrogen-acceptor (O−H···A) angles, which are related to the hydrogen bonds in the crystal structure of xylitol.

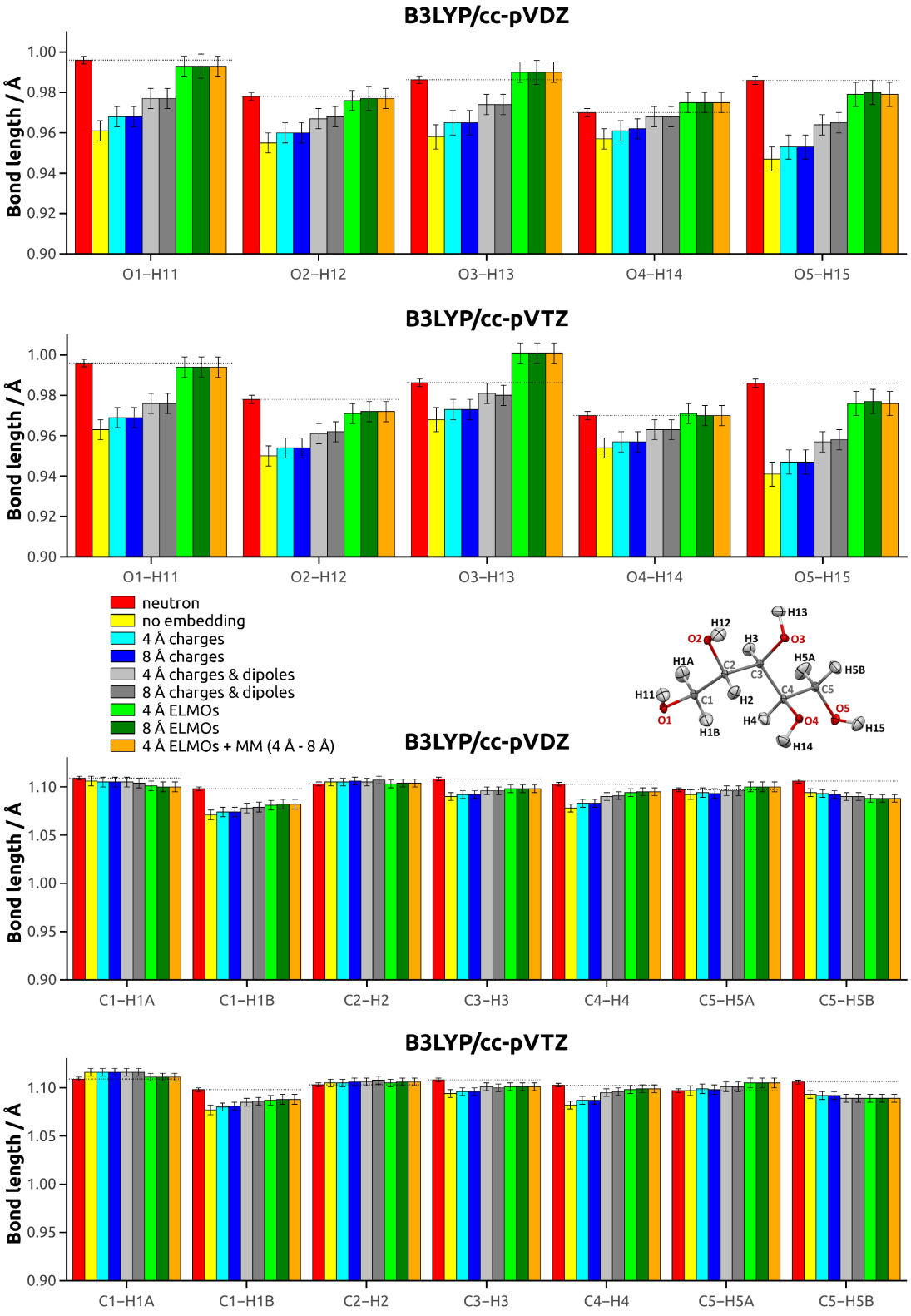


Figure 5.4: O−H (top) and C−H (bottom) bond lengths obtained for the different types Hirshfeld atom refinements of xylitol. The corresponding neutron bond distances are also shown together with a picture of the neutron structure[334] indicating the atomic labels. Adapted with permission from reference [329]. Copyright 2021 American Chemical Society.

Let us first focus on the C−H bond distances obtained through HARs that use the basis set cc-pVTZ. In this case, the mean absolute difference is 0.011(8) Å for the traditional HAR without embedding. The discrepancy slightly reduced to 0.010(7) Å when clusters of charges are introduced. The additional use of dipoles further reduces the mean absolute difference to 0.008(5) Å. However, the best agreement with the neutron data is observed for the the bond lengths obtained with the ELMO embedding strategy, for which the mean absolute difference is 0.007(5) Å. Although these results are promising, they are not statistically significant as indicated by the standard deviations.

In contrast, more significant improvements are observed for the O−H bond lengths. For example, for the basis set cc-pVTZ, the mean absolute differences reduce from 0.029(11) Å without embedding to 0.024(10) Å with cluster charges and further to 0.016(8) Å with charges and dipoles. Finally, a much better agreement with the neutron data is obtained with the ELMO or ELMO/MM embedding, for which the mean ratio is 0.006(6) Å. Even larger improvements that are statistically significant were observed for the basis set cc-pVDZ.

For both the C−H and O−H bond lengths, similar trends to the ones described above can be found for the mean ratios, which are also given in Table 5.3. All the mean ratios are below one, which indicates that the HAR bond lengths are on average shorter than the neutron ones. This is consistent with the mean ratios for the H$\cdots$A distances, which are all above one, meaning that they are all longer than the corresponding neutron distances. Nevertheless, the agreement is certainly worse for the HARs without embedding or those with cluster charges (mean ratios of 1.014(7) and 1.012(7) for cc-pVTZ, respectively). The agreement is intermediate for refinements performed with clusters of charges and dipoles (mean ratio of 1.008(6))), while the best agreement is obtained with the ELMO-embedded-HARs strategy. In fact, the corresponding mean ratio drops to 1.001(5). The same trends can also be observed for the mean absolute differences. For the O−H$\cdots$A angles, the lowering in ratios and differences is statistically not significant. However, also in this case, the most accurate results are obtained with HARs based on QM/ELMO and QM/ELMO/MM calculations.

Similar to what has been observed for the individual bond lengths, also from the statistical analysis it is evident that the size of the embedding regions has no or only minor influence on the mean ratios and mean absolute differences compared to the effect of the strategy for describing the crystal environment. Furthermore, from Table 5.3, it is evident that the lowest standard deviations are always associated with the ELMO-embedded-HARs, which means that a lower variability of bond lengths and angles is obtained with this strategy than with the previously used approaches.

Table 5.3: Statistical analysis of distances and angles obtained with the different types of HARs of the xylitol crystal structure. For each distance or angle, mean ratios and mean absolute differences between HAR and neutron values are considered. The first column refers to the value of the quantity, while the second column to the corresponding standard deviation upon averaging. Adapted with permission from reference [329]. Copyright 2021 American Chemical Society.

Mean ratio								
Basis set / type of embedding	C−H		O−H		H···A		O−H···A	
cc-pVDZ								
no embedding	0.989	0.010	0.971	0.010	1.014	0.007	1.002	0.007
4 Å charges	0.990	0.009	0.977	0.009	1.011	0.006	1.001	0.006
8 Å charges	0.990	0.009	0.977	0.009	1.011	0.006	1.001	0.006
4 Å charges & dipoles	0.992	0.007	0.986	0.007	1.007	0.005	1.000	0.005
8 Å charges & dipoles	0.992	0.008	0.986	0.007	1.007	0.005	1.000	0.005
4 Å ELMOs	0.992	0.007	0.999	0.005	1.001	0.003	0.999	0.003
8 Å ELMOs	0.993	0.007	0.999	0.004	1.001	0.002	0.999	0.003
4 Å ELMOs + MM (4 Å - 8 Å)	0.993	0.007	0.999	0.005	1.001	0.002	0.999	0.003
cc-pVTZ								
no embedding	0.992	0.010	0.971	0.011	1.014	0.007	1.003	0.006
4 Å charges	0.994	0.009	0.976	0.010	1.012	0.007	1.003	0.005
8 Å charges	0.994	0.009	0.975	0.010	1.012	0.007	1.003	0.005
4 Å charges & dipoles	0.996	0.008	0.983	0.009	1.008	0.006	1.002	0.003
8 Å charges & dipoles	0.996	0.008	0.984	0.008	1.008	0.005	1.002	0.004
4 Å ELMOs	0.996	0.008	0.999	0.009	1.001	0.005	1.001	0.003
8 Å ELMOs	0.997	0.008	0.999	0.009	1.001	0.005	1.001	0.003
4 Å ELMOs + MM (4 Å - 8 Å)	0.997	0.008	0.999	0.009	1.001	0.005	1.001	0.003

Mean absolute difference								
Basis set / type of embedding	C−H/Å		O−H/Å		H···A/Å		O−H···A/°	
cc-pVDZ								
no embedding	0.013	0.010	0.029	0.014	0.026	0.014	0.9	0.7
4 Å charges	0.012	0.009	0.023	0.012	0.020	0.012	0.7	0.7
8 Å charges	0.012	0.009	0.023	0.012	0.020	0.012	0.7	0.7
4 Å charges & dipoles	0.010	0.007	0.014	0.007	0.013	0.008	0.6	0.4
8 Å charges & dipoles	0.010	0.007	0.014	0.007	0.013	0.008	0.6	0.4
4 Å ELMOs	0.009	0.007	0.004	0.002	0.005	0.002	0.4	0.4
8 Å ELMOs	0.009	0.006	0.004	0.002	0.004	0.002	0.3	0.4
4 Å ELMOs + MM (4 Å - 8 Å)	0.009	0.006	0.004	0.002	0.005	0.002	0.3	0.4
cc-pVTZ								
no embedding	0.011	0.008	0.029	0.011	0.025	0.011	0.8	0.7
4 Å charges	0.010	0.007	0.024	0.010	0.021	0.011	0.7	0.6
8 Å charges	0.010	0.006	0.024	0.010	0.021	0.011	0.7	0.6
4 Å charges & dipoles	0.008	0.005	0.017	0.009	0.014	0.010	0.5	0.4
8 Å charges & dipoles	0.009	0.005	0.016	0.008	0.014	0.009	0.5	0.4
4 Å ELMOs	0.007	0.005	0.006	0.006	0.008	0.004	0.4	0.2
8 Å ELMOs	0.007	0.005	0.006	0.006	0.007	0.004	0.5	0.2
4 Å ELMOs + MM (4 Å - 8 Å)	0.007	0.005	0.006	0.006	0.007	0.004	0.5	0.2

5.3.3 Atomic displacement parameters

As mentioned in Section 2.1.10, hydrogen ADPs obtained from HAR generally seem to be less accurate than the corresponding E−H bond lengths when both are compared to neutron reference values. Nevertheless, as mentioned in the previous chapter (Section 4.3.3), in reference [221] we speculated that an ELMO embedding could maybe improve the hydrogen ADPs resulting from HAR. In the following, this possibility will be evaluated.

For a visual comparison of the ADPs, the refined xylitol structures are shown in Figure 5.5. Except for some differences observed for the ADPs of H13, H14 and H15, all the performed HARs seem to provide very similar results. In comparison to the hydrogen ADPs in the neutron structure, the ones in the HAR structures are sometimes more elongated and differently oriented. For a more quantitative analysis, the mean ratios (Table 5.4) and mean absolute differences (Table 5.5) between HARs and neutron ADPs are also reported. For all hydrogen and non-hydrogen atoms, the differences between the adopted HAR strategies are rather small and in every case significantly smaller than the standard uncertainties. However, some trends can still be found. In particular, for the ADPs of hydrogen atoms bonded to carbon, the mean ratios are quite close to the optimal value of 1.0 for HARs with clusters of charges and dipoles, while all the other strategies provide higher average ratios that are very similar to each other. Furthermore, for the ADPs of hydrogen atoms that are bonded to oxygen, the best mean ratios and lowest mean absolute differences were observed for the HARs with charges and dipoles and for the ELMO-embedded-HARs.

In conclusion, the agreement of HAR and neutron ADPs cannot be substantially improved by the adopted types of embedding and other methods to accurately determine hydrogen ADPs should be considered in the future (compare Section 2.1.10).

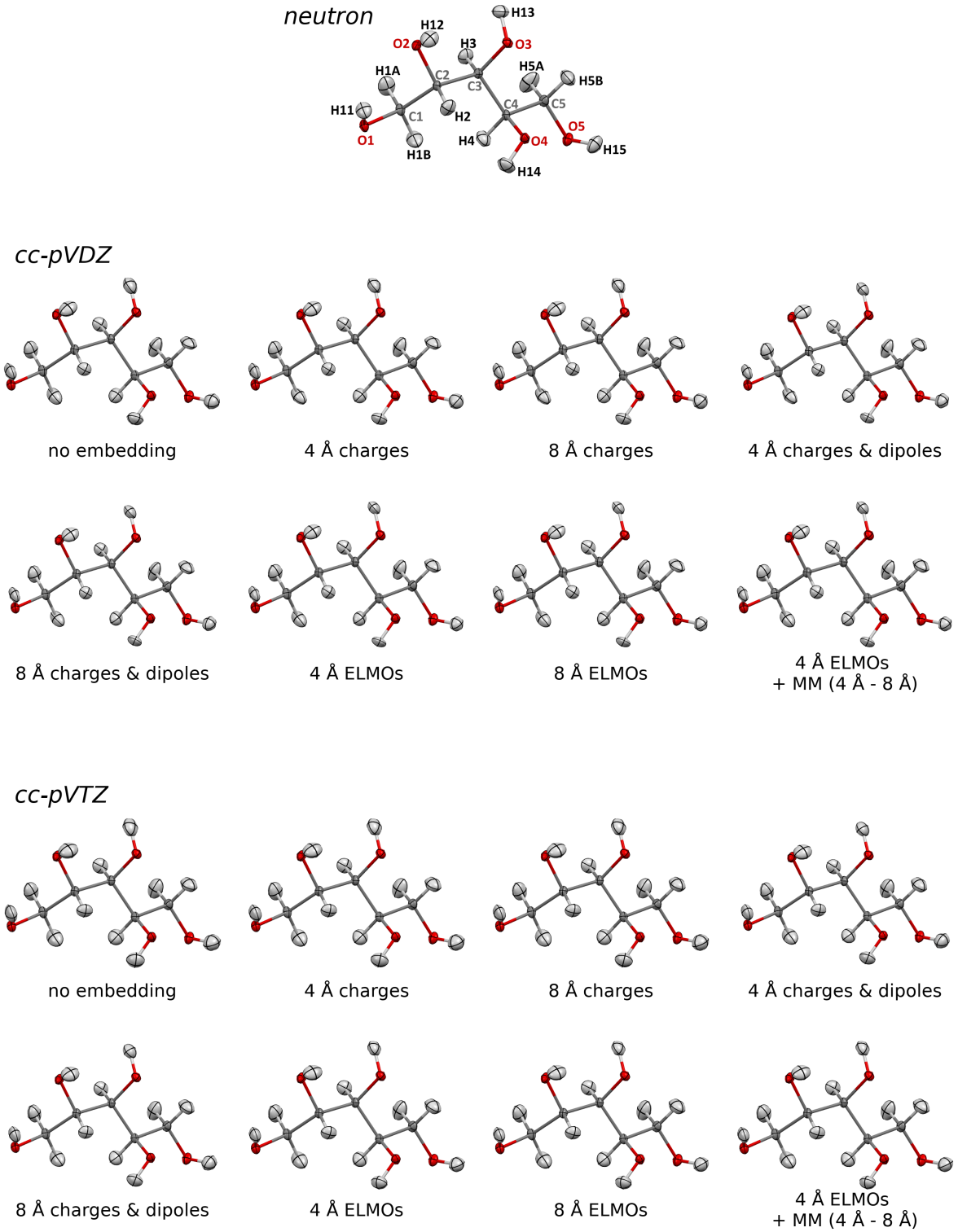


Figure 5.5: First row: Neutron structure from reference [334]. All the rows below: HAR structures of xylitol. All the atomic displacement parameters are depicted at 50% probability level.

Table 5.4: Mean ratios between the HAR and neutron diagonal ADPs $\langle U^{ii}_{\mathrm{HAR}}/U^{ii}_{\mathrm{neutron}}\rangle$ for the different groups of atoms in xylitol. For each group, the first column refers to the value of the quantity and the second column to the corresponding standard deviation upon averaging.

Basis set / type of embedding	Non-H		H bonded to C		H bonded to O	
cc-pVDZ						
no embedding	1.211	0.102	1.047	0.182	1.264	0.339
4 Å charges	1.211	0.101	1.039	0.181	1.182	0.274
8 Å charges	1.211	0.101	1.038	0.181	1.181	0.278
4 Å charges & dipoles	1.211	0.101	1.023	0.183	1.055	0.189
8 Å charges & dipoles	1.211	0.101	1.021	0.184	1.052	0.193
4 Å ELMOs	1.210	0.100	1.040	0.189	0.979	0.174
8 Å ELMOs	1.210	0.100	1.039	0.190	0.976	0.179
4 Å ELMOs + MM (4 Å - 8 Å)	1.210	0.100	1.039	0.190	0.977	0.178
cc-pVTZ						
no embedding	1.207	0.099	1.163	0.202	1.467	0.434
4 Å charges	1.208	0.099	1.153	0.198	1.389	0.349
8 Å charges	1.208	0.099	1.153	0.199	1.388	0.353
4 Å charges & dipoles	1.208	0.098	1.133	0.196	1.256	0.222
8 Å charges & dipoles	1.208	0.098	1.131	0.197	1.253	0.227
4 Å ELMOs	1.207	0.098	1.152	0.205	1.185	0.258
8 Å ELMOs	1.207	0.097	1.150	0.206	1.182	0.263
4 Å ELMOs + MM (4 Å - 8 Å)	1.207	0.098	1.150	0.206	1.183	0.262

Table 5.5: Mean absolute differences between the HAR and neutron ADPs for the different groups of atoms in xylitol. The differences are computed separately for the diagonal elements $\langle|U^{ii}_{\mathrm{HAR}} - U^{ii}_{\mathrm{neutron}}|\rangle$ and the non-diagonal elements $\langle|U^{ij}_{\mathrm{HAR}} - U^{ij}_{\mathrm{neutron}}|\rangle$. For each group of atoms, the first column refers to the value of the quantity and the second column to the corresponding standard deviation upon averaging. Adapted with permission from reference [329]. Copyright 2021 American Chemical Society.

$\langle|U^{ii}_{\mathrm{HAR}} - U^{ii}_{\mathrm{neutron}}|\rangle$/Å^2

Basis set / type of embedding	Non-H		H bonded to C		H bonded to O	
cc-pVDZ						
no embedding	0.0018	0.0007	0.0031	0.0026	0.0055	0.0054
4 Å charges	0.0018	0.0007	0.0030	0.0027	0.0043	0.0041
8 Å charges	0.0018	0.0007	0.0030	0.0027	0.0042	0.0041
4 Å charges & dipoles	0.0018	0.0007	0.0031	0.0028	0.0029	0.0023
8 Å charges & dipoles	0.0018	0.0007	0.0031	0.0028	0.0028	0.0024
4 Å ELMOs	0.0018	0.0007	0.0035	0.0026	0.0028	0.0022
8 Å ELMOs	0.0018	0.0007	0.0035	0.0027	0.0028	0.0022
4 Å ELMOs + MM (4 Å - 8 Å)	0.0018	0.0007	0.0035	0.0026	0.0028	0.0022
cc-pVTZ						
no embedding	0.0018	0.0007	0.0049	0.0031	0.0097	0.0066
4 Å charges	0.0018	0.0007	0.0048	0.0029	0.0082	0.0055
8 Å charges	0.0018	0.0007	0.0048	0.0029	0.0082	0.0055
4 Å charges & dipoles	0.0018	0.0007	0.0047	0.0027	0.0057	0.0038
8 Å charges & dipoles	0.0018	0.0007	0.0047	0.0027	0.0057	0.0038
4 Å ELMOs	0.0018	0.0007	0.0050	0.0028	0.0049	0.0036
8 Å ELMOs	0.0018	0.0007	0.0050	0.0028	0.0048	0.0037
4 Å ELMOs + MM (4 Å - 8 Å)	0.0018	0.0007	0.0050	0.0028	0.0048	0.0037

$\langle|U^{ij}_{\mathrm{HAR}} - U^{ij}_{\mathrm{neutron}}|\rangle$/Å^2 with $i \neq j$

Basis set / type of embedding	Non-H		H bonded to C		H bonded to O	
cc-pVDZ						
no embedding	0.0002	0.0002	0.0035	0.0027	0.0042	0.0027
4 Å charges	0.0002	0.0002	0.0035	0.0027	0.0039	0.0025
8 Å charges	0.0002	0.0002	0.0035	0.0027	0.0039	0.0025
4 Å charges & dipoles	0.0002	0.0002	0.0035	0.0027	0.0036	0.0022
8 Å charges & dipoles	0.0002	0.0002	0.0035	0.0027	0.0036	0.0022
4 Å ELMOs	0.0002	0.0002	0.0035	0.0028	0.0036	0.0021
8 Å ELMOs	0.0002	0.0002	0.0035	0.0028	0.0036	0.0021
4 Å ELMOs + MM (4 Å - 8 Å)	0.0002	0.0002	0.0035	0.0028	0.0036	0.0021
cc-pVTZ						
no embedding	0.0002	0.0002	0.0031	0.0019	0.0042	0.0038
4 Å charges	0.0002	0.0002	0.0030	0.0019	0.0037	0.0035
8 Å charges	0.0002	0.0002	0.0030	0.0019	0.0037	0.0035
4 Å charges & dipoles	0.0002	0.0002	0.0031	0.0018	0.0031	0.0028
8 Å charges & dipoles	0.0002	0.0002	0.0031	0.0018	0.0031	0.0027
4 Å ELMOs	0.0002	0.0002	0.0032	0.0018	0.0032	0.0030
8 Å ELMOs	0.0002	0.0002	0.0032	0.0018	0.0032	0.0030
4 Å ELMOs + MM (4 Å - 8 Å)	0.0002	0.0002	0.0032	0.0018	0.0032	0.0030

5.3.4 Electron densities

To explain the results described above, the electron densities obtained from the QM calculations underlying the different types of HAR were compared. All densities were computed on the IAM structure, which is the common starting point for all HARs in this study. As explained in Section 4.3.4, the largest geometrical changes usually occur at the first HAR iteration. Therefore, the corresponding electron density is crucial for the outcome of the refinement. Moreover, from differences in the electron densities, an explanation for the different results obtained with the different embedding strategies can be found. To compute the difference densities, the B3LYP calculation without embedding was chosen as reference. In other words, the corresponding density was subtracted from all the other densities that were obtained from B3LYP calculations with the different types of embedding. The corresponding electron density differences are shown in Figure 5.6. It is interesting to note that the difference electron densities are mostly caused by the type of embedding, while the size of the surrounding region or the basis set have only a minor influence. This finding is consistent with the previously described trends for bond lengths and ADPs.

For all the adopted embedding strategies, the electron density differences are mainly located on the hydroxy groups of xylitol. The differences are small for the embedding with charges, and increase when also dipoles are considered. The largest differences are obtained for the ELMO-embedded calculations. For the "charges & dipoles" and the ELMO embeddings, charge depletions can be observed for the hydrogen atoms, while charge accumulations can be noticed for the oxygen atoms, particularly in the directions pointing to the surrounding hydrogen atoms. These differences can be associated with the electron density shifts due to the intermolecular hydrogen bonds.

Therefore, from the electron density differences, it can be concluded that the ELMO embeddings and to a lower extent also the embeddings with charges and dipoles account for the changes in the electron density due to the crystal environment. For this reason, O−H bond lengths that are in much better agreement with the neutron reference values can be obtained using the new ELMO-embedded-HAR strategy.

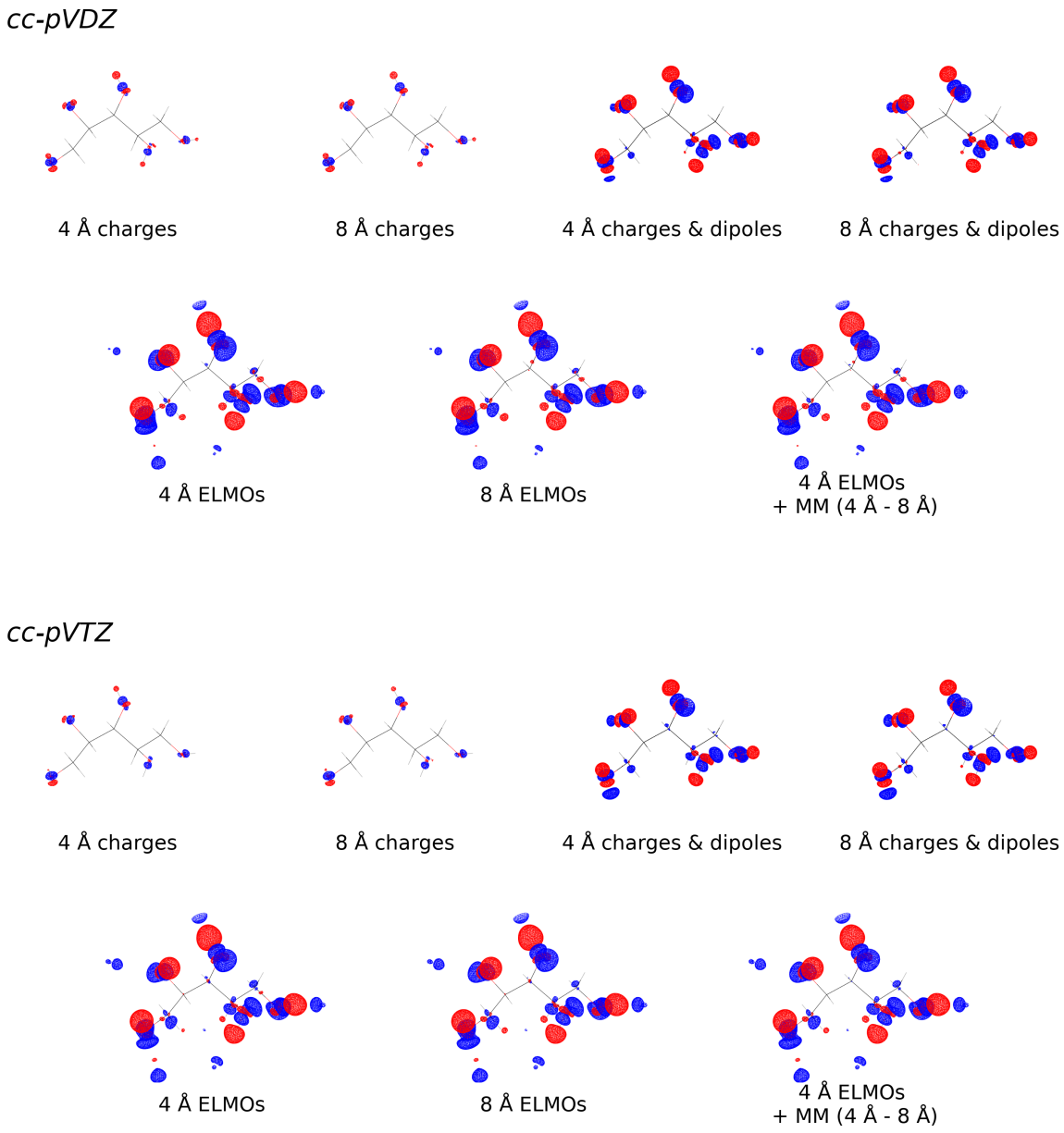


Figure 5.6: Electron density differences for xylitol obtained for the different embedding schemes compared to the calculation without embedding. All the electron densities were computed on the IAM structure. Contour level: 0.003 e bohr^{-3}; Colors: blue (positive) and red (negative). The orientation of the xylitol molecule is the same as in Figure 5.5. Adapted with permission from reference [329]. Copyright 2021 American Chemical Society.

5.4 Conclusions and outlook

In summary, in this chapter, a new strategy to account for the crystal environment in HAR has been described. The new technique, which is based on the QM/ELMO and QM/ELMO/MM embedding strategies, provides E−H bond lengths that are in excellent agreement with the neutron reference values, also for bonds that are involved in intermolecular hydrogen bonds. From the test refinements of the xylitol structure it can be concluded that calculations with a 4 Å ELMO embedding in combination with cc-pVDZ are sufficient for obtaining reliable results. An outer molecular mechanics layer or larger basis sets are not required. However, further testing of the ELMO embedding method is necessary in the future before more general recommendations can be given.

Nevertheless, the direct comparison of the different types of embedding shows that at least a calculation with charges and dipoles is necessary to obtain E−H bond lengths that are closer to the neutron benchmark values. However, the best agreement is almost always obtained when the environment is described at ELMO level. For this reason, this type of embedding should always be considered for the refinement of crystal structures that form strong intermolecular hydrogen bonds with the surroundings of the molecular reference unit.

In DFT, the concept of Jacob's ladder[31] is used to rank different types of exchange-correlation functionals (for details compare Section 1.1.3). Borrowing this concept, a similar ladder could also be established for X-ray structure refinement techniques (Figure 5.7). On the first rung we could clearly locate the IAM and the "heaven" would be represented by the true crystal structure. The new ELMO-embedded-HAR technique could thus be seen as a new rung on the ladder of structure refinement. In the future, also other fully QM embedding methods (Section 1.4.2) could be coupled to HAR, with the goal of obtaining more and more accurate crystal structures through the refinement of X-ray data.

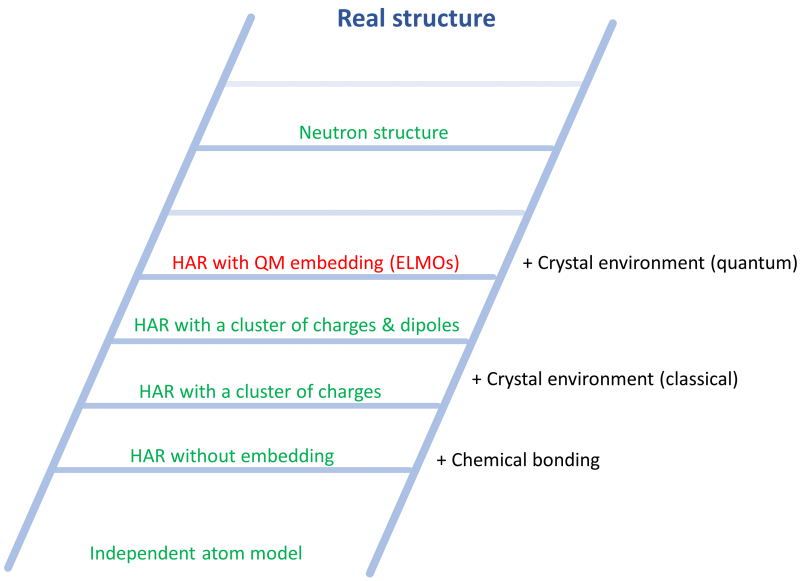


Figure 5.7: Jacob's ladder for the refinement of hydrogen atoms.

6 HAR-QM/ELMO for refinement of organometallic compounds

6.1 Introduction

So far, all the compounds investigated in the context of this thesis were organic molecules or larger biomolecules, in which hydrogen atoms are typically of protic character and are involved in bonds of the type $E^{\delta-}$–$H^{\delta+}$. However, despite their rather simple electronic structure, hydrogen atoms have remarkably flexible bonding abilities.[335] Hence, depending on the bonded element, hydrogen atoms can also be hydridic and are thus involved in bonds of the type $E^{\delta+}$–$H^{\delta-}$. Especially transition metal hydrides have been studied extensively, because of their important role in catalysis and because they may be used as hydrogen storage materials.[336–338] Interestingly, such hydrides also occur in the active sites of hydrogenase enzymes that catalyze the uptake or evolution of hydrogen.[339,340] However, also hydridic hydrogen atoms bonded to main group elements deserve attention, since they can be used as reducing agents or as precursors for other metal hydrides.[335]

However, locating these hydrogen atoms accurately by means of X-ray diffraction experiments is particularly challenging. In general, hydrogen atoms scatter X-rays very weakly, and only up to a resolution of approximately 1.5 Å.[341] Moreover, when hydrogen atoms are bonded to heavy elements, the small maxima in the residual density corresponding to the hydrogen atoms can be obliterated by the so-called Fourier truncation ripples. These ripples occur due to the termination of the Fourier series.[303,341] For these reasons, locating hydrogen atoms next to heavy atoms is particularly difficult.

From the experimental point of view, it is crucial to obtain a very complete X-ray dataset, where both the low resolution reflections[255] and the high resolution ones[242] were carefully measured. The former are important because they carry the information about the hydrogen atoms, while the latter reduce the Fourier truncation ripples, which become weaker and are located in closer proximity to the heavy element if the dataset is measured up to high resolution.[341]

From the software development point of view, in the case of HAR, a particular difficulty used to be that the Hirshfeld atoms in *TONTO* had not been optimized for elements heavier than krypton.[303] This difficulty has been recently overcome with the development of *NoSpherA2*,[303] which allowed to accurately determine the position of hydrogen atoms even when they are bonded to heavy metal atoms.[262]

As already explained in Section 2.3.4, HAR is based on iterative wavefunction calculations. Therefore, it is necessary to consider several aspects related to the presence of heavy elements. For example, relativistic effects should be considered.[303,325–327,342] Furthermore, metalorganic compounds often include bulky ligands with many atoms. Therefore, the iterative wavefunction calculations are either very time consuming or even unfeasible if post-HF methods are used.

To overcome this limitation, the QM/ELMO strategy offers appealing features as post-HF calculations can be performed for a specific part of a system at a significantly reduced computational cost.[206] In this chapter, a procedure will be proposed to refine the structures of compounds with hydrogen atoms bonded to heavy elements using the novel HAR-QM/ELMO method. The central idea at the basis of this technique is the following: the QM/ELMO method will be coupled with HAR by extending the *NoSpherA2* interface. In particular, the heavy element-hydrogen bond of interest (and possibly some surrounding atoms) will be treated at a high QM level (e.g. CCSD), whereas the bulky ligands will be described by means of frozen ELMOs. The goal of this study will be to evaluate whether the use of post-HF methods could improve the description of heavy elements in HARs, and whether they could also improve the heavy element-hydrogen bond lengths. However, before carrying out this study, the HAR-ELMO and HAR-QM/ELMO methods need to be implemented in *NoSpherA2*.

Since this work is currently still ongoing, only preliminary developments and envisaged refinements will be described in this chapter. In particular, in Section 6.2, the implementation of HAR-ELMO in *NoSpherA2* will be described, followed by a short description of the envisaged implementation and validation of the HAR-QM/ELMO technique in Section 6.3.

6.2 HAR-ELMO in *NoSpherA2*

Before implementing the HAR-QM/ELMO refinement technique in *NoSpherA2*, as a first step, we have added the option to perform HAR-ELMO refinements. This development occupied an important part of my research activity in the last year of my Ph.D. and will be the main topic of this section.

NoSpherA2[303] is a recently developed interface to perform HARs within the *olex2* GUI. As explained in Section 2.3.4, any Hirshfeld atom refinement consists in the following steps: (i) a single-point QM calculation to obtain the electron density, (ii) the partitioning of the electron density into Hirshfeld atoms, (iii) the computation of structure factors, (iv) a least-squares refinement, (v) a check for convergence. These steps are performed iteratively until the refinement parameters converge. In the previous chapters, the different types of refinements were performed using either *lamaGOET* or our in-house script, which both interface a QM software with *TONTO*. In *NoSpherA2*, all five steps can be carried out independently of *TONTO*. In particular, the user can choose between different QM programs (e.g. *ORCA*, *Gaussian*) for the single-point calculations (step i). The partitioning of the electron density and the computation of structure factors is performed by *NoSpherA2* itself (steps ii and iii), while *olex.refine* is used to carry out the least-square refinement (step iv).

Since *olex.refine* uses the same input instructions as *SHELXL*, the new HAR refinements are very user friendly because they require the same input files as standard IAM refinements. Moreover, the new interface overcomes many technical limitations. For example, disorder refinement is available in *olex.refine*. This is a crucial development in the context of protein refinements with HAR, because the lack of this possibility used to be one of the two main limitations for such refinements, as described in Section 3.5. This is the reason why we have also implemented HAR-ELMO refinements in *NoSpherA2*.

To accomplish this task, the *ELMOdb* software was added to the programs that are already interfaced in *NoSpherA2* for computing the electron density at the first step of each HAR cycle. Hence, it is not necessary anymore to always perform a single-point calculation at each refinement cycle. Instead, if the corresponding ELMOs are available, the wavefunction can be obtained through transfers of these ELMOs.

As a usual HAR, also HAR-ELMO refinements start from an initial IAM refinement. Furthermore, in the case of HAR-ELMO refinements, it is currently necessary to provide an initial PDB file in AMBER format for the system under exam. In the future, we envisage to write the PDB file automatically in *olex2*. However, this option is not yet implemented. Nevertheless, it is only necessary to provide a PDB file once at the beginning of the HAR-ELMO refinement, since the coordinates of the atoms are then automatically updated in the course of the refinement.

The *olex2* GUI (see Figure 6.1) provides different options for HAR-ELMO refinements. These will be described in the following.

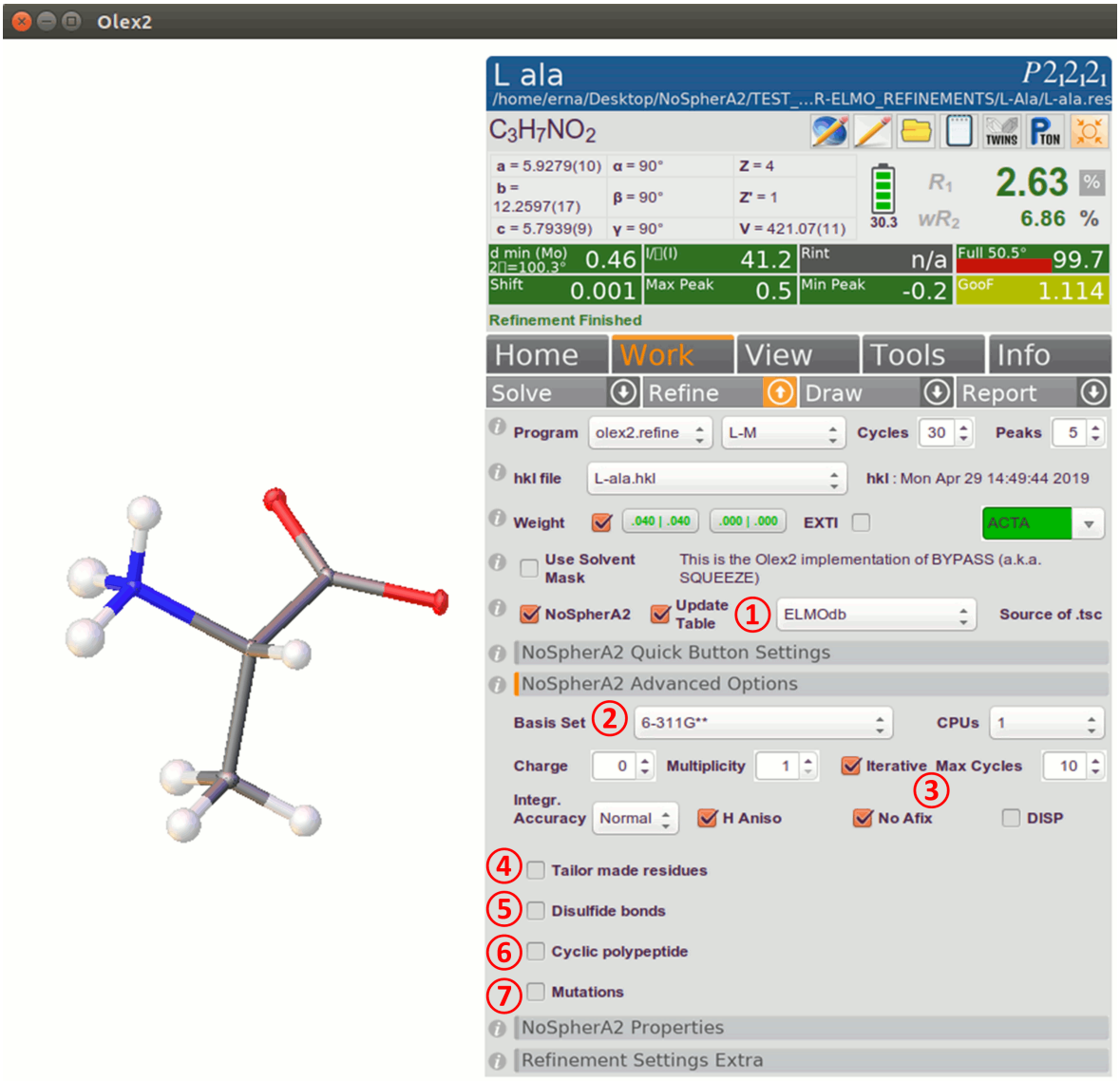


Figure 6.1: Setup of a HAR-ELMO refinement in the *olex2* GUI.

In the simplest case, all required ELMOs are available in the database for all the residues of the system under exam. Starting from an initial IAM refinement, the user can immediately continue with a HAR-ELMO refinement. In the *olex2* GUI, it is only necessary to enable the "NoSpherA2" option and to select "ELMOdb" from the list of available QM programs in the dropdown menu (point (1) in Figure 6.1). In the "NoSpherA2 Advanced Options" block, one needs to select a basis set from the five standard basis sets available in the ELMO libraries (point (2) in Figure 6.1). This is enough to carry out one HAR-ELMO cycle with *NoSpherA2*. To perform iterative cycles of steps (i) to (v), one also needs to tick the "iterative" box in the "NoSpherA2 Advanced Options" (point (3) in Figure 6.1).

In a more complicated case, the system under exam contains residues with ELMOs that are not available in the database. In this case, the input for tailor-made residues can be provided within the *olex2* GUI (point (4) in Figure 6.1). If the system contains disulfide bridges, it is necessary to specify the number of such bonds and the labels of the involved residues in the *ELMOdb* input (point (5) in Figure 6.1) because the program needs to know to which position the ELMOs associated with the disulfide bond need to be rotated. Similarly, if the system contains cyclic polypeptide chains, the number of chains and the corresponding starting and terminal residues need to be given in the *ELMOdb* input (point (6) in Figure 6.1). In this way, the ELMOs corresponding to the peptide bond between the starting and terminal residue of the cyclic polypeptide are rotated correctly by the *ELMOdb* program.

As mentioned above, one of the most important advantages associated with the implementation of the HAR-ELMO technique in *NoSpherA2* is that disorder refinements are now possible. In practice, this is done by transferring the ELMOs separately to each conformer and by also computing the atomic scattering factors separately. Afterwards, they are properly combined for all the conformations through the *NoSpherA2* interface. The same procedure is also applied if single-point QM calculations are performed instead of ELMO transfers. However, for HAR-ELMO refinements, it is necessary to differentiate two particular cases. In the first case, the residues are the same in all conformations, whereas in the second case, some of the residues differ (for an example see below). In the first case, the ELMOs can be transferred as described above since the same PDB file can be used for all the present conformations. However, in the second case, different PDB files need to be provided for each mutated confomer. This option is enabled by selecting the "mutations" option in the *olex2* GUI (point (7) in Figure 6.1).

6.2.1 Test compounds

The previously described software developments are currently being tested in cooperation with the group of Dr. Simon Grabowsky (University of Bern). For example, to test the option for disorder refinement, we envisage to refine the structure of the cyclic polypeptide shown in Figure 6.2. The polypeptide crystallized with two butanol and one water solvent molecule out of which one of the butanol molecules is disordered.[343] Additionally, we are currently refining the mixed *L*-asparagine·H_2O/ *L*-aspartic acid crystal structure,[344] which we use as test case for the "mutations" option (see above). A preliminary structure refined with HAR-ELMO is shown in Figure 6.3. Moreover, as already mentioned in Chapter 3, we also plan to re-refine the structure of crambin with a proper treatment of the disordered residues.

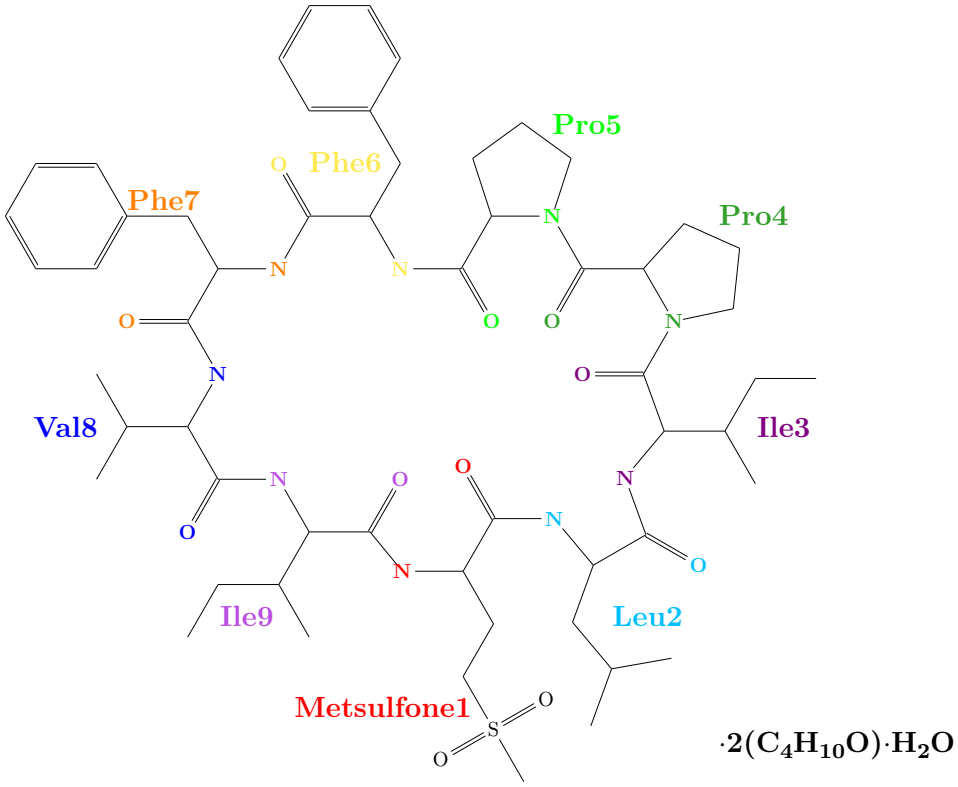
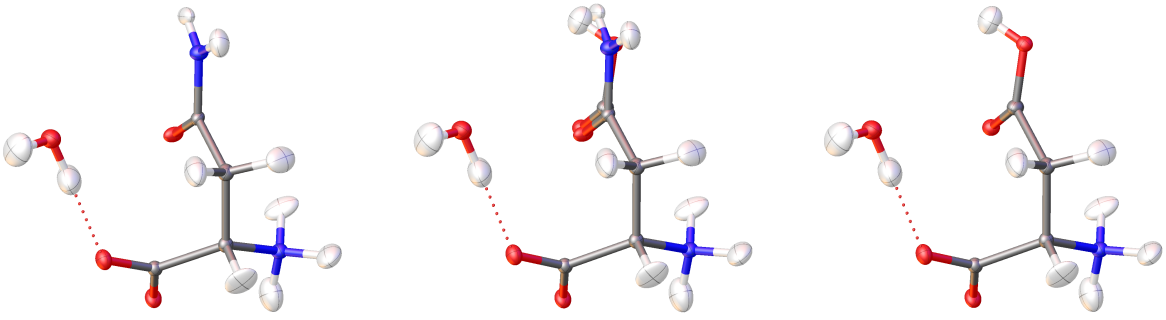


Figure 6.2: Structure of the cyclic polypeptide for which HAR-ELMO refinement in *NoSpherA2* is envisaged. More precisely, the structure is a butanol-water solvate of the cyclolinopeptide *cyclo*(Metsulfone1-Leu2–Ile3–Pro4–Pro5–Phe6–Phe7–Val8–Ile9) with all amino acids in *L*-configuration. In the crystal structure[343] (CSD identifier: AYUQIV), one of the butanol solvent molecules is disordered.



(a) Part 1 (b) Part 1 and 2 (c) Part 2

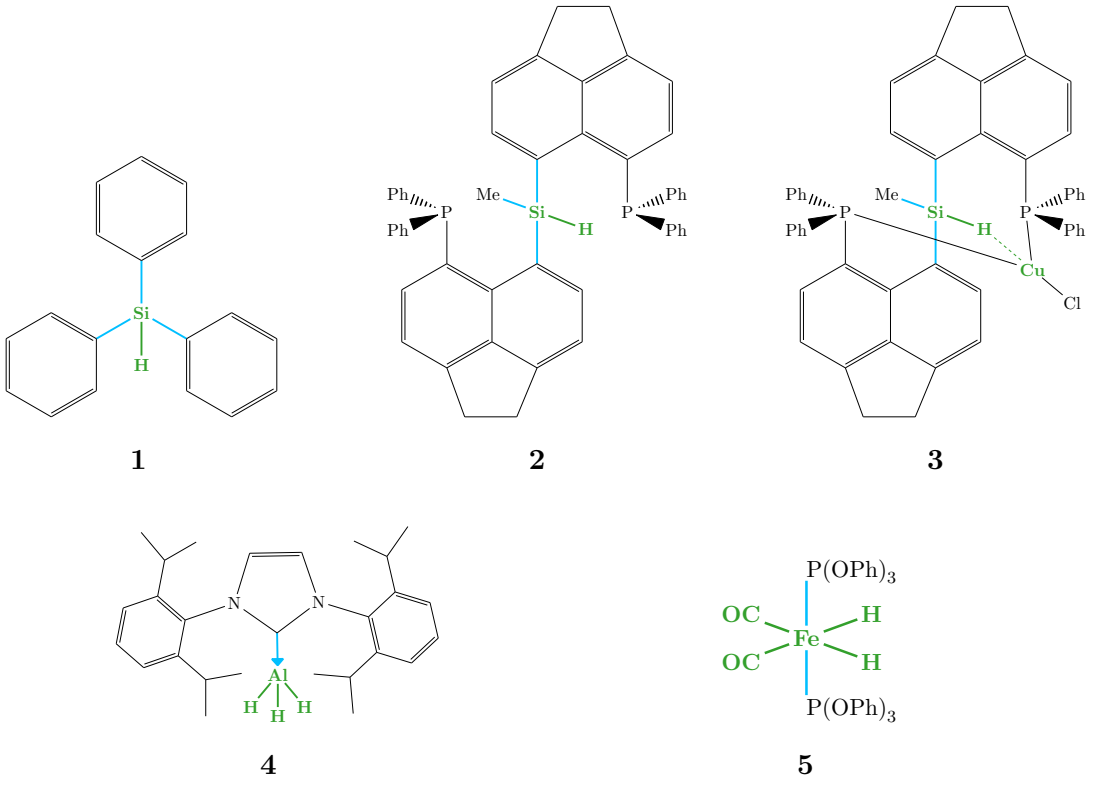
Figure 6.3: Preliminary HAR-ELMO refinement of the mixed *L*-asparagine·H_2O/ *L*-aspartic acid crystal structure[344] in *NoSpherA2*. The panels (a) and (c) show the two disordered parts, while panel (b) shows the overlayed parts.

6.3 HAR-QM/ELMO in *NoSpherA2*

Starting from the previously described developments, we are currently also implementing the HAR-QM/ELMO technique in the *NoSpherA2* interface. As described in the introduction, this strategy could be particularly interesting for the refinement of compounds in which a hydrogen atom is directly bonded to a heavy element (E−H bonds). In particular, we aim at evaluating the influence of post-HF methods on the refinement results, which can be seen as a follow-up study of the investigation described in Chapter 4. For example, it would be interesting to see if the use of these methods has an influence on the lengths of the different E−H bonds or on the residual densities of the heavy atoms E.

6.3.1 Test compounds

For validating the HAR-QM/ELMO technique, we envisage to refine the structures of the compounds depicted in Figure 6.4.



1 **2** **3**

4 **5**

Figure 6.4: Compounds for testing HAR-QM/ELMO refinements in *NoSpherA2*.

The structures **1-3** represent a set of compounds in which the central Si−H bond is involved in different types of interactions, resulting in significantly different Si−H bond lengths. The three compounds have been previously characterized by Hupf *et al.*, who have carried out neutron experiments of all three structures.[345] Summarizing the results obtained by Hupf and coworkers, in triphenylsilane (compound **1**), the Si−H bond is 1.494(6) Å long and the negatively polarized hydrogen atom is involved in a short intermolecular interaction with a positively polarized hydrogen atom in the phenyl ring of a neighboring triphenylsilane molecule in the crystal structure. In contrast, in compound **2**, the Si−H bond is only 1.484(7) Å long

and the hydrogen atom is not involved in intermolecular interactions. Compound **2** represents the starting material for compound **3**, in which the Si−H bond is 1.509(7) Å long and is involved in an agostic interaction with the Cu^+ ion. In addition to the neutron structures, IAM refinements of the corresponding X-ray crystal structures have been also published by Hupf *et al.* (CSD codes: TPHSIL02, QUBJUV01, QUBKAC).

An X-ray structure of compound **4** has been described in the literature[346] (CSD code: LUFVUD) with a monoclinic crystal system. During my master thesis[347] at the university of Bremen (group of Dr. Simon Grabowsky), we measured a high resolution (0.65 Å) X-ray dataset and a neutron dataset of a triclinic polymorph of compound **4**. Both measurements are currently still unpublished. The asymmetric unit of the triclinic polymorph consists of two molecules. Therefore, six symmetry independent Al−H bonds are present in the crystal structure and have lengths of 1.566(15) Å, 1.574(13) Å, 1.591(13) Å, 1.598(11) Å, 1.606(17) Å and 1.616(14) Å. Hence, they span a range of 0.05 Å and it could be interesting to evaluate whether the HAR-QM/ELMO technique is able to yield accurate bond lengths or at least provide the same trends as the neutron reference.

Finally, compound **5** represents a transition metal hydride. Iron is the heaviest element considered in this series. For compound **5**, both X-ray[348] and neutron[349] datasets are available in the literature (CSD codes: QOSZON and QOSZON01, respectively) and its structure has been previously refined with HAR by Woińska and coworkers.[262] In the crystal structure, the two Fe−H bonds are symmetry independent. In the neutron structure they are 1.529(2) and 1.521(2) Å long, while the HAR refinements performed by Woińska *et al.* yielded bond lengths of 1.522(15) and 1.505(14) Å, respectively. Woińska and coworkers also concluded that the refinements could be further improved by testing more sophisticated methods like MP2 or CCSD for the HARs. In fact, this is the aim of the current study.

The selected compounds should enable us to thoroughly test the HAR-QM/ELMO technique. In fact, for all five compounds, a neutron reference is available for comparing the E−H bond lengths resulting from HAR. Furthermore, the present E−H bonds differ significantly in their lengths. Finally, the structures of the compounds contain quite large ligands, which makes them also realistic test cases for HAR-QM/ELMO refinements because post-HF calculations of the complete molecules are unfeasible.

6.3.2 Envisaged procedure

Before the HAR-QM/ELMO refinements can be performed, it is necessary to implement the option to obtain the wavefunction from QM/ELMO calculations in *NoSpherA2*. In the *olex2* GUI, the user should be able to choose the orbitals for the QM region as well as the QM method and the basis set. In the backend of *NoSpherA2*, only minor changes are required because single-point calculations with *Gaussian09* can be performed already. Therefore, it is only necessary to modify the writing routine for the input file.

In the QM/ELMO calculations, the orbitals associated with the green atoms (E, H) as well as the green (E−H) and blue bonds (E−C or E−P) in Figure 6.4 will be part of the QM region, while the ligands (black atoms and bonds) will be described with transferred ELMOs. The small size of the QM region should make post-HF calculations feasible for these molecules, even at CCSD level. To test the influence of post-HF methods on the E−H bond lengths and

the residual densities, it would be interesting to perform HAR-QM/ELMO refinements at HF/ELMO and at CCSD/ELMO levels for all compounds, and compare the resulting E−H bond lengths and the residual densities in the QM region.

6.4 Perspectives

In this chapter, recent developments concerning the implementation of the HAR-ELMO and HAR-QM/ELMO refinement techniques in *NoSpherA2* have been described. Furthermore, the envisaged procedure for testing and applying these refinement strategies has been outlined. While the HAR-ELMO method will be used to refine disordered structures of different polypeptides and proteins, the HAR-QM/ELMO technique will be exploited to further evaluate the necessity of using post-HF methods for HARs. In particular, we are currently refining the structures of compounds that include Al−H, Si−H and Fe−H bonds with the HAR-QM/ELMO technique, with the goal of investigating whether these bond lengths or the residual densities of the aluminum, silicon or iron atoms improve by using post-HF methods.

Furthermore, the implementation of HAR-ELMO and HAR-QM/ELMO in *NoSpherA2* represents a further and crucial step forward for applying HAR to protein structures because disorder refinement is now possible. Moreover, the HAR-QM/ELMO technique will offer the possibility to refine structures of systems in which a highly accurate description of one particular part of the molecule is required. For example, this includes compounds similar to the ones shown in Figure 6.4, but also host-guest assemblies or metalloproteins.

7 Summary and conclusions of Part I

The refinement of X-ray structures is of paramount importance to obtain three dimensional structures of small molecules as well as proteins and other large systems. However, it is sometimes forgotten that the majority of X-ray refinements is based on a crude approximation. In fact, the underlying IAM approximates the electron density as a sum of independent spherical atomic densities. Therefore, deformations of the electron density due to chemical bonding are completely neglected. This has important repercussions also on the obtained structure because E−H bond lengths are systematically too short. This limitation can be overcome by using aspherical models in the refinement. A detailed introduction to standard X-ray and neutron structure refinements as well as aspherical refinement techniques was given in Chapter 2 of this dissertation.

One particular aspherical refinement technique is the Hirshfeld atom refinement method, which is based on QM calculations to obtain the electron density. By exploiting HAR, accurate E−H bond lengths can be obtained using only X-ray data. The work presented in this part of the thesis focused on different ways of computing the electron densities for Hirshfeld atom refinements. We aimed at speeding up the QM calculations to increase the maximum size of the systems that can be refined with HAR. However, we also tested different approaches to further increase the accuracy of the results obtained through HAR.

In Chapter 3, the development of the HAR-ELMO technique was shortly described. It aims at addressing the need of faster QM calculations for HAR and at the application of the technique to protein refinements. As the results of the validation test showed, the HAR-ELMO approach provides E−H bond lengths that are in very good agreement with those resulting from traditional HARs. However, the HAR-ELMO results were obtained at a significantly reduced computational cost. Therefore, the HAR-ELMO technique can at least in principle be applied to significantly larger structures than the original HAR strategy. In Chapter 3, this was demonstrated on refinements of two polypeptides and a small protein. Nevertheless, at the time of the study, these refinements were only possible due to the very good quality of the considered datasets and thanks to the introduction of quite strong approximations. In fact, usual protein structures are generally measured up to resolutions that are too low for applying HAR or HAR-ELMO refinements. Moreover, it was not possible to properly refine disordered structures at the time of the study. This second limitation has only very recently been overcome with the development of the *NoSpherA2* interface in *olex2* (see also Chapter 6).

As mentioned above, a second goal of the work presented in this thesis was to increase the accuracy of the results that can be obtained using HAR. In a first study, we tested the possibility of improving the results using post-HF methods. This study is the main topic of Chapter 4. In particular, the high-resolution and low-temperature structure of *L*-alanine was refined with HAR using six different QM methods (ELMOs, HF, BLYP, B3LYP, MP2 and CCSD) in combination with three different basis sets (def2-SVP, def2-TZVP, and def2-

TZVPP). Despite some influences of the different QM levels on the E−H bond lengths and anisotropic displacement parameters of hydrogen atoms could be observed, the corresponding differences were much smaller than the standard deviations associated with these parameters. Therefore, for the refinement of *L*-alanine, post-HF methods for HAR are not necessary. However, different results could be obtained for other types of compounds. This will be further investigated in future studies, as described in Chapter 6.

One particular shortcoming of the investigations described in Chapters 3 and 4 and also of other HAR studies is that some of the obtained E−H bond lengths are still systematically shorter than those resulting from refinements of neutron data. This deviation is mostly observed for polar E−H bond lengths that are involved in strong intermolecular interactions with the neighboring molecules in the crystal. Even when the QM calculations are performed with an embedding of charges and dipoles, these bond lengths are still shorter than the neutron reference values. As an alternative to the traditional embedding of charges and dipoles, we coupled the HAR and QM/ELMO techniques giving rise to the ELMO-embedded HAR strategy, which is the main topic of Chapter 5. In this new approach, the central molecular reference unit is described using a QM method, while the crystal environment is also explicitly taken into account in the calculation because ELMOs are transferred to all molecules that are within a certain radius around the molecular reference unit. To test the ELMO-embedded HAR strategy, we performed refinements of xylitol exploiting different types of embeddings. All the lengths of the present polar O−H bonds improved significantly by using an embedding. In particular, the new ELMO-embedded HAR strategy provided O−H bond lengths that are either in optimal agreement or at least systematically closer to the neutron reference values than the bond lengths resulting from HARs without embedding or with classical embeddings using clusters of charges and dipoles.

In the last chapter of this first part, in Chapter 6, the newest developments for implementing the HAR-ELMO and HAR-QM/ELMO techniques in the *NoSpherA2* interface of *olex2* were described. This implementation will overcome many previous limitations. For example, it will allow refinements with disorder treatment and restraints. In Chapter 6, the envisaged strategies for testing the new implementations were also described. The HAR-ELMO technique will be applied to properly refine disordered structures of amino acids, polypeptides and a small protein. Moreover, the HAR-QM/ELMO approach will be exploited to refine organometallic hydrides. In particular, the heavy element and the bonded hydrogen will be described at post-HF level, while the bulky ligands will be treated with frozen ELMOs. The goal of the study is to improve the description of the metal atom, obtain a better accuracy and precision of the E−H bond length and further clarify whether post-HF techniques are useful for HARs.

In conclusion, the work presented in this first part of the thesis showed that (i) HAR can become significantly less computationally expensive by combining it with the ELMO libraries, even allowing the refinement of protein structures; (ii) post-HF methods are probably not necessary for improving HARs of small organic molecular crystal structures; (iii) more accurate polar E−H bond lengths can be obtained from HAR using an ELMO embedding; (iv) user-friendly ways of performing HARs could become reality even for challenging molecules in the near future.

Possible future directions of research certainly include further testing and validation of the HAR-ELMO, HAR-QM/ELMO and ELMO-embedded HAR techniques. For example, the third approach should be applied to other compounds than xylitol. In general, for future validation of new HAR variants and similar refinement strategies, it would be very useful to assemble a well-balanced database of X-ray structures. Ideally, this database should include very high-resolution X-ray datasets, which could be also used for benchmarking different QM methods (as suggested in Section 4.4). However, datasets of lower resolution should also be added because they could help to estimate which resolution would be required for applying certain techniques. Finally, at least for some of the structures, neutron datasets measured at the same temperature as the X-ray ones should also be available for comparing structural parameters of hydrogen atoms.

The database suggested above would probably include mostly small molecular crystal structures but also HARs of large molecules should certainly not be dismissed. However, before this could become a routine application, further developments will be necessary. For example, datasets of protein crystals should ideally be measured up to the maximum resolution that can be reached. Furthermore, the software used for HAR is currently optimized for the refinement of small molecular structures. For this reason, these programs could either be further extended or coupled with existing protein refinement and visualization software. Nevertheless, a combined effort from both experimental and development points of view could make HAR a very useful tool also for protein refinements.

Part II

Analysis of non-covalent interactions in polypeptides and proteins

8 Introduction to the analysis of non-covalent interactions

Non-covalent interactions are ubiquitous in nature and play a key role in biochemistry. A long list of different interaction types has been described in the literature, including hydrogen bonds,[350,351] halogen bonds (and the analogous tetrel, pnictogenic and chalcogen bonds),[352] interactions involving aromatic rings[353,354] (such as π-π-stacking,[355] cation$\cdots\pi$[356,357] or anion$\cdots\pi$[358,359] interactions), and many others.

The importance of these interactions can hardly be overestimated. For example, they are responsible for the existence of the liquid phase, the unique properties of water,[360,361] the packing of molecules in crystal structures,[362] as well as the structures and functions of biomolecules such as DNA and proteins.[360,361] Especially in large biosystems, many different types of non-covalent interactions are acting together, stabilizing the structures of these systems, and allowing for their biological activity. For example, the double helical structure of DNA exists because of a complicated interplay of covalent and non-covalent interactions, such as in-plane hydrogen bonds and electrostatic interactions and out-of-plane dispersion interactions, as well as interactions with the solvent molecules.[361,363] Likewise, several types of non-covalent interactions contribute to the stability of protein structures[364] and are involved in the folding of proteins,[365–367] in molecular recognition processes[368–371] as well as in protein-protein interactions.[372–374]

Compared to covalent bonds, non-covalent interactions are significantly weaker, and can be reversibly formed and broken. Hence, non-covalent interactions are crucial for nature not despite but due to their weakness.[360,368] However, since they are characterized only by small changes in the energy and the electronic structure, their accurate description by means of theoretical approaches is far from being trivial.[360,375]

Nevertheless, to accomplish this task, different techniques have been developed. They could be grouped into methods based on geometrical criteria, interaction energies or real space quantities. The aim of this chapter is to give a short introduction to the different types of strategies. However, the main focus in this part of the thesis is on two techniques that fully belong to the real space indicators and that will be described in details in the introduction sections of the next two chapters.

8.1 Analysis based on geometries

In the analysis of the structures obtained from crystallographic refinement or molecular dynamics (MD) simulations, non-covalent interactions are often identified using distance-based criteria.[222,376–378] Furthermore, a definition of suitable geometrical criteria can also be useful to estimate how often a specific interaction occurs in a large number of systems, as for example done to (initially) find out the number of cation$\cdots\pi$, anion$\cdots\pi$ and n-π^* interactions in proteins[357,359,379,380] or to identify metal-environment contacts in metalloproteins.[381] How-

ever, these predictions can only be as good as the underlying structures.[382]

Another option, which is commonly applied in crystallography, is to evaluate whether the distance between two atoms is smaller or larger than the sum of the corresponding van der Waals radii.[383–387] If it is smaller, the atoms are considered to be interacting. However, in this context, the van der Waals radii obtained by Bondi are often used despite they were not devised for this application as pointed out by Bondi himself.[386] Therefore, newer definitions of van der Waals radii should be considered.[387,388]

In summary, the analysis based on geometrical criteria represents an easy and computationally inexpensive approach to approximately detect non-covalent interactions. However, less directional non-covalent interaction as van der Waals interactions cannot be reduced to simple contacts between atoms. Therefore, geometrical criteria can rather help to establish expectations about the presence of some types of non-covalent interactions, but do not provide general information about interaction energies or changes in the electron density associated with the formation of non-covalent interactions.

8.2 Analysis based on energies

8.2.1 Interaction energies from force fields

For certain types of interactions, the interaction energies can be accessed through force fields. However, while electrostatic and van der Waals interactions can be quite accurately described using this approach, the energies associated with hydrogen bonding are systematically underestimated.[389] Furthermore, for host-guest complexes, the interaction energies obtained through quantum mechanical calculations or with generalized force fields were found to differ significantly.[390] Therefore, force fields are only of limited use for determining the interaction energies of non-covalent interactions, also because they are lacking information about quantum effects like the transfer of protons and electrons.[391]

8.2.2 The supermolecular approach

A promising alternative that still keeps a relatively low computational cost is represented by semiempirical methods. In fact, while traditional techniques of this kind fail to accurately describe non-covalent interactions, more recently introduced correction terms for dispersion and hydrogen bonding yield interaction energies that are well comparable with fully QM calculations.[392] Similar results have been reported for DFT, where current functionals do not automatically include dispersion.[361,393,394] Therefore, interaction energies obtained from DFT have to be taken with caution[361] unless correction terms for dispersion are added. In fact, these corrections, as for example Grimme's dispersion correction (DFT-D3),[395] lead to significant improvements in the resulting interaction energies.[361,393,394] Nevertheless, the benchmark techniques to compute interaction energies remain correlated wavefunction methods. In this regard, especially CCSD(T) is nowadays established as the benchmark technique for computing interaction energies for small systems.[360,361,391,396]

To practically compute interaction energies, all the above-described semiempirical and traditional QM techniques exploit the supermolecular approach, which consists in calculating

the energy difference between the complex (AB) and the sum of all the relevant monomer (A and B) energies:

$$E_{\text{int}}(AB) = E(AB) - (E(A) + E(B)) \tag{8.1}$$

Since the basis set in the calculation of the individual energies is finite, the so-called basis set superposition error[397] needs to be accounted for. This error occurs because the energy of the complex (E^{AB}) is computed in the basis of both monomers, while the energies of each individual monomer (E^{A} and E^{B}) are computed using only the basis functions centered on the atoms of that monomer. This leads to an artificial stabilization of the complex compared to the dissociated monomers. To avoid this error, also the energies of the monomers need to be computed in the complete basis set of the complex. Therefore, Equation (8.1) is modified in the following way:

$$E_{\text{int}}(AB) = E(AB) - (E(A^{AB}) + E(B^{AB})), \tag{8.2}$$

where the superscript indicates that the calculation is performed in the basis set of the complete dimer.[360]

8.2.3 Variational energy decomposition analysis

To gain chemical insights into the interaction energies, an energy decomposition analysis (EDA)[398,399] can be performed. In this regard, variational and perturbation based approaches have been developed. In this subsection, the focus is on the former type, whereas a perturbation based approach will be described in the next subsection.

The central idea of the variational EDA approaches is to decompose the interaction energy of the system AB into different components, where each term is associated with a particular type of interaction between the monomers A and B. To compute the different energy components, intermediate wavefunctions are constructed at each step of the analysis, and the energy differences between these wavefunctions are computed. Common components of the interaction energy include electrostatic, polarization, exchange, correlation and charge transfer contributions. However, in practice, many different ways to decompose the interaction energy exist, resulting in a variety of variational EDA schemes, in which the exact definition of the individual energy terms may differ.[398]

8.2.4 The symmetry adapted perturbation theory

An alternative to the previously described supermolecular approach and to the variational EDA strategies is represented by the perturbation methods. The main idea of the perturbation approach is to add correction terms for the interactions to the Hamiltonian of the non-interacting system.[400] One particular pertubation method is the symmetry adapted perturbation theory (SAPT),[401,402] which provides accurate interaction energies and, at the same time, offers the possibility for an energy decomposition analysis, in which all the energy components have a physical meaning.[400] In particular, terms for electrostatic, induction, dispersion and exchange contributions are obtained in SAPT.[402]

The advantage of computing the total interaction energy with SAPT is that this energy can be computed directly i.e. without the need for subtracting large absolute energy values

as in the supermolecular approach.[361] Furthermore, SAPT provides interaction energies that are as accurate as those resulting from supermolecular approach based on CCSD(T) calculations, and avoids the problem of correcting for the basis set superposition error. However, also the associated computational cost of a SAPT computation is as high as the one for CCSD(T) calculations (compare Section 1.1.4). In this regard, a substantial improvement can be achieved by combining SAPT with DFT in the so-called SAPT-DFT,[403–405] which is significantly faster than the original SAPT technique and less sensitive to the choice of the functional than the supermolecular DFT approach. However, the basis set superposition error needs to be corrected again in SAPT-DFT.[360,361]

8.3 Analysis based on real space indicators

While interaction energies are important to quantify interactions, real space methods can identify regions in space that are associated with covalent and non-covalent interactions. Hence, they are particularly useful tools offering an intuitive and easy interpretation of non-covalent interactions.[400] Different types of real space indicators exist.

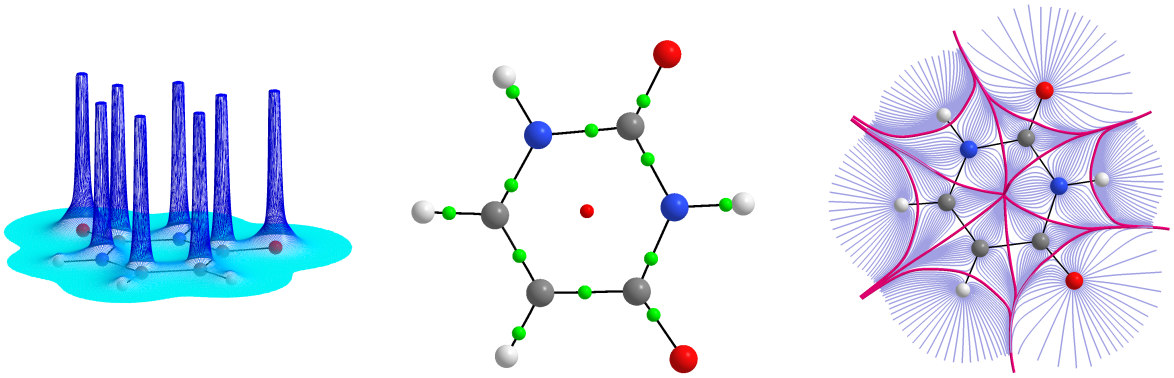
One possibility is to use the electrostatic potential (ESP) that measures the electrostatic energy that a positive charge would experience at different points in space. A positive value of the ESP indicates a repulsion of the charge, whereas a negative value corresponds to an attractive interaction.[406] The ESP is often mapped on an isosurface of the electron density and can give valuable insights into non-covalent interactions.[400,407–409] However, these maps can be misinterpreted because changes in the ESP are often associated with changes of the local electron density, which is not necessarily the case as pointed out by Wheeler and Houk.[406]

Another possibility is to use functions that measure the localization of electrons in a certain region of space, as done for example in the electron localization function (ELF)[410] or ELI-D[318,411] strategies. Both techniques measure the probability of finding electrons with the same spin in a certain region of space and can provide a very intuitive interpretation of chemical bonding. However, for non-covalent interactions, the applicability of ELF is limited to strong electrostatic interactions like hydrogen bonds.[400,412]

Finally, in the context of real space indicators, the electron density itself can be also interpreted together with its derived scalar and vector fields. In this regard, a popular and well-established tool for the analysis of chemical bonding is based on the quantum theory of atoms in molecules (QTAIM).[413] Furthermore, specialized techniques were particularly developed for the analysis of non-covalent interactions. Examples include the NCI method,[319,375,414] which is simultaneously based on the electron density and the reduced density gradient, and the IGM technique,[321,415] which is also based on the electron density and a so-called δg-descriptor. They offer the possibility to identify, classify and to some extent also quantify non-covalent interactions. Both techniques play a major role in the work presented in this part of the thesis and will be further introduced below and in the next chapters. However, before that, the main ideas of QTAIM will be summarized below because some basic concepts of this approach are also used in NCI and IGM.

8.3.1 Quantum theory of atoms in molecules

The popular QTAIM[413,416] approach is based on the topological analysis of the three dimensional electron density function $\rho(\boldsymbol{r})$. As an example, the relief map of the electron density in the plane of uracil is shown in Figure 8.1a. As can be observed from this figure, the electron density is dominated by maxima at the positions of the carbon, nitrogen and oxygen nuclei. Peaks associated with hydrogen atoms are much smaller. However, also in between the nuclear positions, the electron density is not featureless. In fact, minima and saddle points exist in between the maxima and can give valuable insights into the interactions between the atoms. Therefore, one of the major aspects of a QTAIM analysis is to identify all the critical points in a molecule by means of a topological analysis of its electron density.



(a) Relief map (b) Molecular graph (c) Gradient map

Figure 8.1: (a) Relief map of the electron density in the molecular plane of uracil. Values were truncated at 226 e bohr^{-3}. (b) Molecular graph of uracil. Bond paths are shown in black, bond critical points in green and the ring critical point in red. (c) Gradient map in the molecular plane of uracil. Trajectories of the gradient $\nabla\rho(\boldsymbol{r})$ are shown in violet, while zero-flux surfaces around each atomic basin are depicted in purple.

The key[417] for analyzing the topology of the electron density is the corresponding gradient vector, which is given by

$$\nabla\rho(\boldsymbol{r}) = \begin{pmatrix} \frac{\partial\rho(\boldsymbol{r})}{\partial x} \\ \frac{\partial\rho(\boldsymbol{r})}{\partial y} \\ \frac{\partial\rho(\boldsymbol{r})}{\partial z} \end{pmatrix} \tag{8.3}$$

Those points where the gradient is zero ($\nabla\rho(\boldsymbol{r}) = 0$) are critical points of the electron density. These can be classified according to the second derivatives of the electron density. For three dimensional functions (like the electron density), nine second partial derivatives need to be computed, which can be arranged in the 3×3 Hessian matrix:

$$H(\boldsymbol{r}) = \begin{pmatrix} \frac{\partial^2\rho(\boldsymbol{r})}{\partial x^2} & \frac{\partial^2\rho(\boldsymbol{r})}{\partial x\partial y} & \frac{\partial^2\rho(\boldsymbol{r})}{\partial x\partial z} \\ \frac{\partial^2\rho(\boldsymbol{r})}{\partial y\partial x} & \frac{\partial^2\rho(\boldsymbol{r})}{\partial y^2} & \frac{\partial^2\rho(\boldsymbol{r})}{\partial y\partial z} \\ \frac{\partial^2\rho(\boldsymbol{r})}{\partial z\partial x} & \frac{\partial^2\rho(\boldsymbol{r})}{\partial z\partial y} & \frac{\partial^2\rho(\boldsymbol{r})}{\partial z^2} \end{pmatrix} \tag{8.4}$$

This matrix can be diagonalized and, depending on the sign of the three eigenvalues λ_1, λ_2 and λ_3 at a critical point, four different types of critical points can be identified. If all three eigenvalues are negative, the point corresponds to a local maximum in all three dimensions.

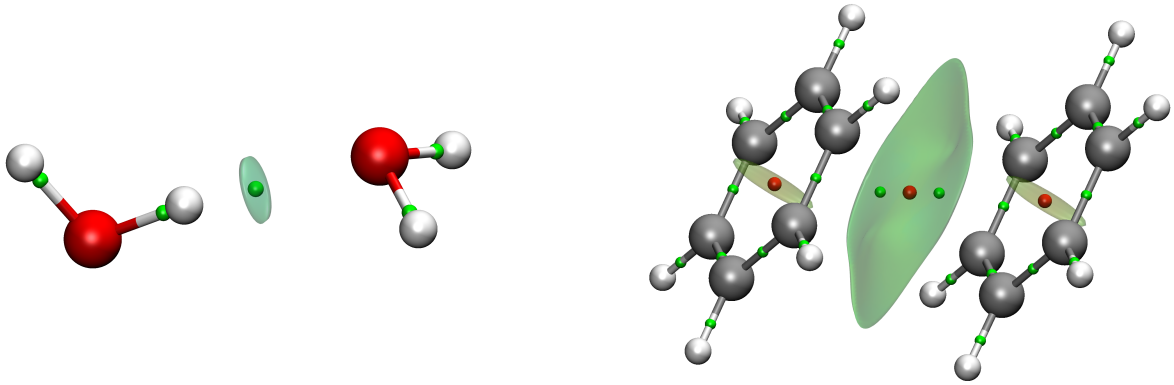
These maxima are called nuclear attractors and usually correspond to the positions of nuclei. A path of maximum electron density between two nuclear attractors is called bond path. The network of bond paths in a molecule is called molecular graph, which is shown in Figure 8.1b for uracil. At a particular point on the bond path, the electron density reaches a minimum, which is called bond critical point. From the perspective of the other two directions that are perpendicular to the bond path, the value of the electron density reaches a maximum at the bond critical point. In Figure 8.1b, the bond critical points are shown in green color. In addition to nuclear attractors and bond critical points, two other types of critical points can be found. One of them is called ring critical point, which is depicted in red color in the middle of the uracil ring in Figure 8.1b. At the ring critical point the electron density reaches a minimum inside the ring plane, while the point corresponds to an electron density maximum in the direction perpendicular to the ring plane. In analogy to the nuclear attractors, also minima in all three dimensions may exist (all three eigenvalues λ_1, λ_2 and λ_3 are negative). These points are called cage critical points.

To further characterize and classify bonds in QTAIM, certain bond properties are calculated at the bond critical point. For example, the value of the electron density at the bond critical point can indicate the strength of the interaction. In fact, at the bond critical point, the electron density usually takes values above 0.20 a.u. in covalent bonds and values below 0.10 a.u. in closed-shell interactions (such as hydrogen bonds, van der Waals interactions etc.).[416] Furthermore, bond critical points can be classified according to the laplacian of the electron density, which is defined as:

$$\nabla^2\rho(\boldsymbol{r}) = \frac{\partial^2\rho(\boldsymbol{r})}{\partial x^2} + \frac{\partial^2\rho(\boldsymbol{r})}{\partial y^2} + \frac{\partial^2\rho(\boldsymbol{r})}{\partial z^2} = \lambda_1 + \lambda_2 + \lambda_3 \tag{8.5}$$

In general, a negative value of the laplacian indicates a charge accumulation, while a positive value corresponds to a charge deplecion. Therefore, a negative laplacian at the bond critical point is a sign for covalent bonding, whereas a positive lapacian indicates a closed-shell interaction.

However, it is important to note that the presence of a bond critical point and a bond path can indicate bonding between atoms,[418] but the bond paths should not be confused with chemical bonds in the sense of the Lewis picture.[419] In fact, bond paths and bond critical points rather indicate the presence of interatomic interactions, which can be both covalent and non-covalent.[400] For example, in the QTAIM analyses of the water and benzene dimers (compare Figure 8.2) the bond critical points are located in between covalently and non-covalently bonded atoms. Therefore, QTAIM is a valuable tool for analyzing these interactions. However, for some types of non-covalent interactions, the description offered by QTAIM is not very intuitive.[400,420] For example, in the benzene dimer (Figure 8.2b), one would expect the π-π interaction to be rather delocalized, as indicated by the green isosurface. However such a delocalized picture is not obtained in the QTAIM analysis. In fact, two bond critical points and one ring critical point are observed in between the benzene monomers, providing a too localized and hard to interpret picture of this non-covalent interaction. To overcome this limitation, methods like the NCI technique have been developed. In fact, the green isosurfaces in Figure 8.2 are obtained using the NCI method, which will be described



(a) H_2O dimer (b) Benzene dimer

Figure 8.2: QTAIM analysis of the (a) water and (b) benzene dimers. Bond and ring critical points are shown as green and red balls, respectively. The transparent isosurfaces are obtained from NCI analyses, and reveal non-covalent interactions. For details about NCI analyses see Chapter 9.

in detail in the next chapter.

Other than the analysis in terms of critical points, QTAIM also offers a natural definition of atoms. In this regard, it is crucial to note that the gradient vector given by Equation (8.3) has a direction. Hence, also the corresponding gradient paths (trajectories) have a direction. Therefore, all the nuclei in a molecule are attractors of the gradient and they are the points where the trajectories are ending (see, for example, the trajectories in the plane of uracil depicted in Figure 8.1c). In any molecule, there exist certain surfaces that are not crossed by any of the trajectories of the gradient field. These surfaces are called zero-flux surfaces (see again Figure 8.1c). Each point $\boldsymbol{r}$ on these surfaces satisfies the condition

$$\nabla\rho(\boldsymbol{r}) \cdot \boldsymbol{n}(\boldsymbol{r}) = 0 \tag{8.6}$$

where $\boldsymbol{n}(\boldsymbol{r})$ is the unit vector normal to the zero-flux surface. These surfaces naturally partition the molecule into "atomic basins", which contain all the trajectories belonging to the basin. In QTAIM, an atom in a molecule is defined by its nuclear attractor and its basin. Different atomic properties are accessible through this definition, for example the volumes and charges of atoms.

Furthermore, based on the definition of atoms in QTAIM, the interacting quantum atom (IQA) technique[421–423] offers the possibility for an energy decomposition analysis. In fact, in IQA, the total energy of the system is partitioned into a sum of intra-atomic and inter-atomic energy contributions. The interaction energies for two groups of atoms can then be computed by summing all the interatomic interaction energies between the atoms that belong to different groups.[424] IQA thus allows the study of covalent, non-covalent and steric interactions[425] and it is fair to say that this technique can be considered as the ultimate benchmark for quantitatively analyzing these interactions using real space indicators.[322] However, an IQA analysis is computationally very demanding, which prevents its application to large systems.[322,400]

8.3.2 Specialized tools for analyzing non-covalent interactions

For the reasons described in the previous subsection, namely the non-intuitive picture for certain non-covalent interactions provided by QTAIM and the very high computational cost of methods like IQA, there is a demand for specialized tools to rapidly and reliably identify non-covalent interactions also in large systems.

The most prominent technique developed in this context is probably the non-covalent interaction (NCI) method,[319,375,414] which is based on the analysis of the electron density and the reduced density gradient in the real space. This technique will be the main subject of Chapter 9. Furthermore, based on the ideas of the NCI method, also other techniques have been developed. One example is the independent gradient model (IGM),[321,415] which is also based on the electron density and the definition of a non-interacting reference system. More details about the IGM technique will be given in Chapter 10.

8.4 Outlook to the next chapters

Non-covalent interactions play a major role in defining structures and functions of biological macromolecules. However, the analysis of non-covalent interactions is difficult because of the size of the systems and because the interactions are characterized by very subtle changes in the energy or density. In fact, the existing methods are either not very accurate (like the geometrical criteria or force fields) or computationally too expensive (like the supermolecular approach based on CCSD(T) calculations, SAPT or IQA) for applications to large systems.

Nevertheless, to identify, classify and possibly also rank non-covalent interactions in large systems, the NCI and IGM techniques can be considered as reasonable compromises between the approaches mentioned in the previous paragraph. However, both NCI and IGM crucially depend on the calculation of the electron density. Of course, traditional QM calculations may be used to compute this density, but, as previously pointed out, they cannot be applied to large molecules with a reasonable computational cost (compare Section 1.1.4). Therefore, when applied to large systems, both the NCI and IGM techniques resorted to the promolecular approximation, where the molecular electron density is approximated as the sum of spherically averaged atomic densities that were calculated on isolated and often also neutral atoms. However, this approximation has not only noteworthy limitations in the context of crystallography (see the previous part of this thesis), but, as we will see in this part of the thesis, it also leads to biased results in the framework of non-covalent interaction analyses.

To overcome the drawbacks associated with the use of the promolecular approximation, both the NCI and the IGM have been recently coupled with the ELMO libraries, giving rise to the corresponding NCI-ELMO[320] and IGM-ELMO[322] strategies, which will be the main topics of the next two chapters.

9 The NCI-ELMO technique

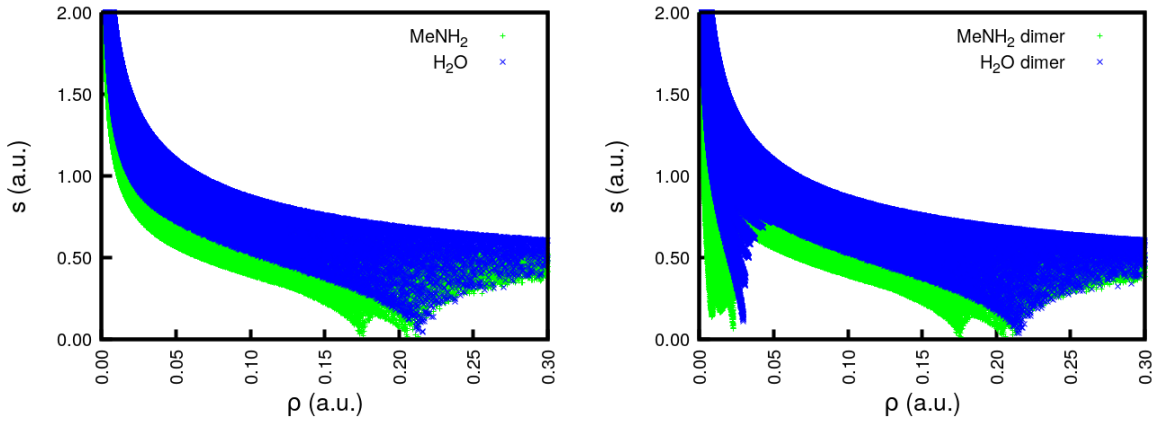
9.1 Introduction to the NCI technique

The NCI index[319,375,414,420,426] is a topological approach that is based on the electron density and the reduced density gradient (RDG). The latter is given by the following expression:

$$s(\boldsymbol{r}) = \frac{1}{C_S} \frac{|\nabla\rho(\boldsymbol{r})|}{\rho(\boldsymbol{r})^{4/3}} \tag{9.1}$$

where $C_S = 2(3\pi^2)^{1/3}$. In particular, the RDG is used in the framework of DFT to correct the deficiencies of the uniform electron gas. In practice, it measures the deviation from a homogeneous electron distribution. In fact, it takes very large positive values in regions that are far from the molecule, where the density decays to zero exponentially. In contrast, it assumes very small values (approaching to zero) in regions that can be associated with covalent and non-covalent interactions. The detection of the latter is the aim of the NCI method.[319]

To accomplish this task, the RDG is plotted against the electron density (in so-called two dimensional (2D) NCI plots).[319] In Figure 9.1a, the 2D NCI plots are shown for the single molecules of methanol and water. In both situations, the RDG approaches zero for density values above 0.15 a.u., which corresponds to the covalent bonds in the two molecules. The same spikes are also observed in Figure 9.1b where the plots for the water and methylamine dimers are shown. However, for the dimers, additional troughs are also obtained for density values below 0.05 a.u. In the NCI method, these troughs are signatures of non-covalent interactions between the dimers.



(a) Monomers (b) Dimers

Figure 9.1: Plots of the RDG against the electron density for water and methylamine, in panel (a) for the single molecules and in panel (b) for the homodimers of the two molecules.

In general, non-covalent interactions correspond to those peaks in the 2D NCI plots,

where the density and the RDG are both low, while covalent bonding is associated with peaks where the density is higher. Therefore, covalent and non-covalent interactions can be easily distinguished within the NCI method. However, the different types of non-covalent interactions (like hydrogen bonds, van der Waals interactions or steric crowding) appear in the same regions of the 2D NCI plots.[319] This can be seen in Figure 9.2a, where the RDG is plotted against the electron density for a dimer of acetic acid and uracil. For clarity, the higher density values (corresponding to covalent interactions) have been neglected.

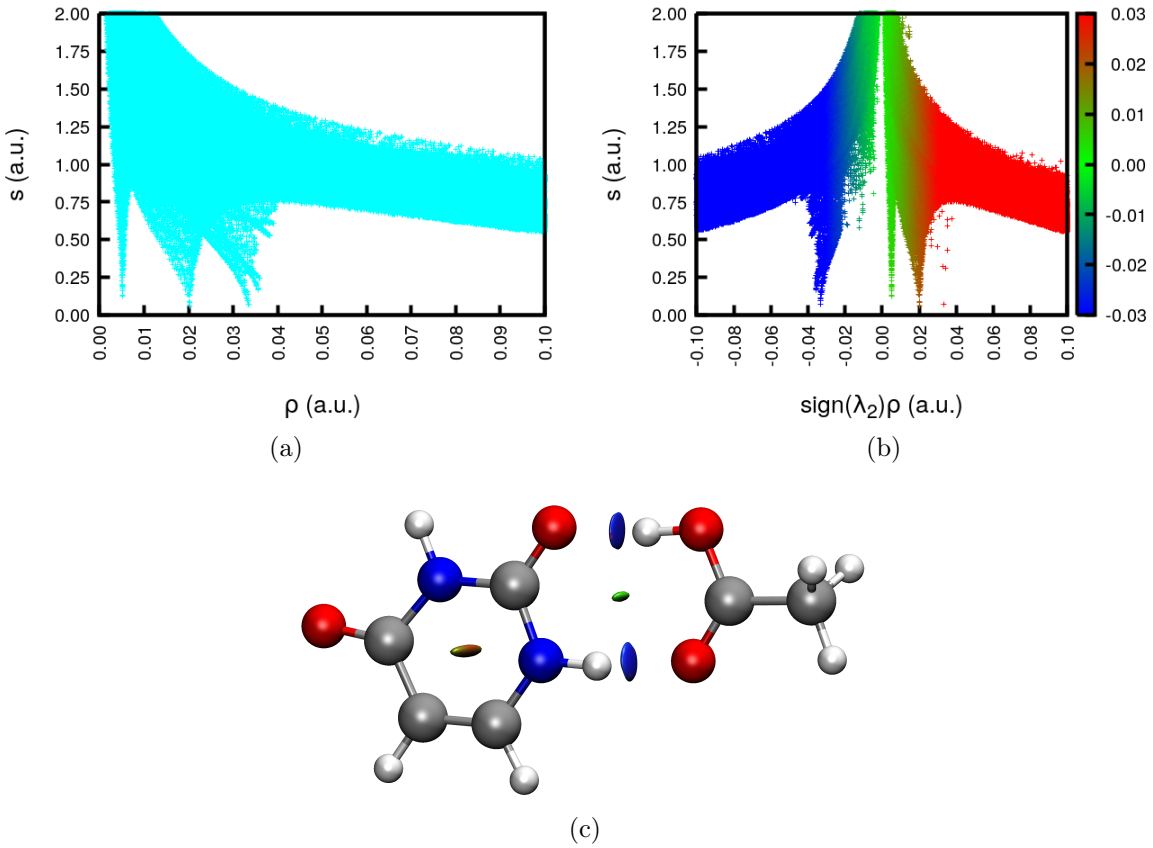


Figure 9.2: NCI plots for the dimer formed by acetic acid and uracil: (a) the RDG is plotted against the electron density (in this case, different types of non-covalent interactions overlap); (b) the RDG is plotted against the "signed" electron density and the different types of interactions can be distinguished (2D NCI plot); (c) corresponding isosurfaces of the RDG, which allow us to locate the different types of interaction in the real space (3D NCI plot).

To further distinguish the different types of interactions, the value of the electron density is multiplied by the sign of the second eigenvalue of the electron density Hessian λ_2 (compare Equation (8.4)). In fact, depending on the type of interaction, λ_2 can either take a positive or negative value. In the case of bonding interactions such as hydrogen bonds, λ_2 is negative. In contrast, it is positive in situations with steric crowding, where several atoms interact without being bonded. Finally, for van der Waals interactions, λ_2 can be either positive or negative. However, the corresponding peaks generally appear in lower density regions in the 2D NCI plots. In fact, Figure 9.2b shows essentially the same plot as Figure 9.2a, but this time the electron density values are multiplied by the sign of λ_2. Additionally, a color scheme is applied, where blue corresponds to negative values, green to intermediate values and red to positive values of the electron density multiplied by λ_2 (henceforth called the "signed"

electron density).

Furthermore, to identify the non-covalent interactions in the 3D space, isosurfaces of the RDG are depicted in 3D NCI plots and enclose the regions associated with the non-covalent interactions. The same color scheme that is used for the 2D plots is also applied to the 3D NCI plots. For example, the RDG isosurfaces for the dimer of acetic acid and uracil are shown in Figure 9.2c.

In NCI analyses, the 2D and 3D plots are interpreted together. For the example of the dimer of acetic acid and uracil both plots (Figures 9.2b and 9.2c) indicate that all three types of non-covalent interactions (hydrogen bonds, van der Waals interactions and steric crowding) are present. The O−H$\cdots$O and N−H$\cdots$O hydrogen bonds between the monomers correspond to blue disked shaped RDG isosurfaces in Figure 9.2c. In the 2D NCI plot, these interactions appear as troughs in the negative region of the "signed" electron density. Within the uracil ring, the red isosurface in the 3D NCI plot is associated with steric crowding, for which the corresponding peak is located in the positive region of the "signed" electron density in Figure 9.2b. Finally, the green isosurface in Figure 9.2c indicates a van der Waals interaction and corresponds to the narrow green peak in the middle of Figure 9.2b.

9.1.1 Types of electron densities in the NCI analyses

For an analysis of non-covalent interactions through the NCI index, the only required input is an electron density for the system under investigation. It is not necessary to know the interactions before performing such an analysis and also the RDG can be calculated if the electron density is known. For the NCI analysis, the underlying electron densities can be obtained either from experiments or from theoretical calculations.[375] For example, experimental electron densities may result from multipole model refinements[237,280,281,427,428] or X-ray constrained wavefunction fittings[122–129]. However, both methods require an excellent quality of the measured X-ray data and are currently not applicable to large biological systems.

Therefore, in this chapter, densities obtained from theoretical calculations will be used. For small systems, the NCI analysis is ideally[420] based on densities resulting from quantum chemical calculations. It has been shown[429] that the influence of the different QM methods (e.g. HF, DFT, MP2, CCSD) on the NCI results is very small. Moreover, calculations that do not include dispersion (HF and DFT) can give results that qualitatively agree with those obtained from CCSD computations.[429] However, for larger systems these traditional QM calculations are unfeasible (compare Section 1.1.4).

Therefore, similarly to what is done in crystallography (see Chapter 2), also the NCI analysis often exploits a promolecular density, which is the sum of spherically averaged densities previously calculated on isolated neutral atoms and afterwards stored within the *NCIPLOT* program.[414,426] The advantage of using promolecular densities is that they are fast to compute, but the drawback is that they lack relaxation. When the promolecular NCI results are compared to those obtained for relaxed densities (resulting for example from DFT or CCSD calculations but also from multipole model refinements) non negligible differences can be observed.[428,429] In particular, the peaks in the 2D plots are often shifted to higher values of the electron density. The largest differences are obtained for non-bonded interactions (where $\lambda_2 > 0$). This is also reflected in the 3D plots, since the isosurfaces of the RDG are usually

larger for promolecular densities.[429] As examples, these trends can be observed in Figures 9.3a-d, where the 2D and 3D NCI plots resulting from promolecular and B3LYP densities are shown for the uracil dimer.

Because of these non-negligible differences, it would be highly desirable to use more reliable strategies to compute the densities for the NCI analyses of large systems. However, as already mentioned above and as described in detail in Section 1.1.4, the required QM calculations are impracticable. Therefore, to overcome the limitations of the fully QM calculations on one side and the ones of the promolecular approximation on the other side, the NCI analyses of large systems could benefit from a technique that allows reliable but fast reconstructions of electron densities. To this end, the ELMO libraries were coupled with the *NCIPLOT* program, giving rise to the new NCI-ELMO approach.[320] In Figures 9.3e-f, the 2D and 3D NCI plots obtained with the new NCI-ELMO technique are shown for the uracil dimer, already highlighting the large similarity between the DFT and ELMO results.

The differences between promolecular-NCI, NCI-DFT and the new NCI-ELMO analyses have been thoroughly investigated for different types of non-covalent interactions in polypeptides and proteins.[320] The corresponding study has been published in reference [320] and will be described in detail in Section 9.2.

9.1.2 Quantitative NCI analyses

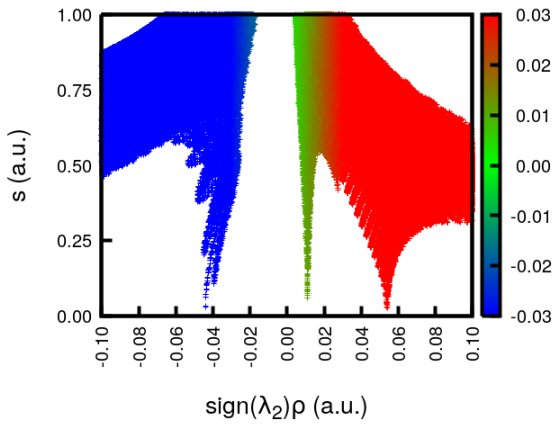
Non-covalent interactions can be identified and classified with the 2D and 3D NCI plots. However, more recently, also a possibility to quantify intermolecular interactions has been introduced within the NCI technique.[414] The aim of this approach is to offer a general and robust method that is computationally less expensive than the strategies described in Chapter 8 and it is based on the parametrization of NCI integrals.

In the following, the current procedure to compute these NCI integrals will be described. The first step is to define an "intermolecular interaction region", i.e. the region to be integrated. To accomplish this task, the total density has to be initially decomposed into contributions of monomers i with corresponding density $\rho_i(\boldsymbol{r})$, so that the total density $\rho(\boldsymbol{r})$ is given by:

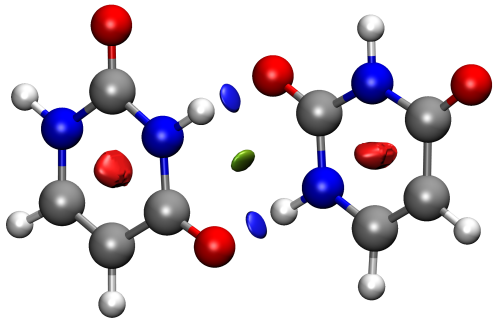
$$\rho(\boldsymbol{r}) = \sum_{i}^{M} \rho_i(\boldsymbol{r}) \tag{9.2}$$

Note that this decomposition is only possible if the method to obtain the density allows for a clear and unambiguous "fragmentation" of the total density into different monomer contributions. Unfortunately, this is not possible for electron densities resulting from traditional QM calculations because the underlying molecular orbitals use the basis functions of the whole system and are thus completely delocalized. Therefore, the procedure to quantify the non-covalent interactions within the NCI approach has been initially developed only for promolecular densities, for which the decomposition according to Equation (9.2) can be unambiguously achieved.

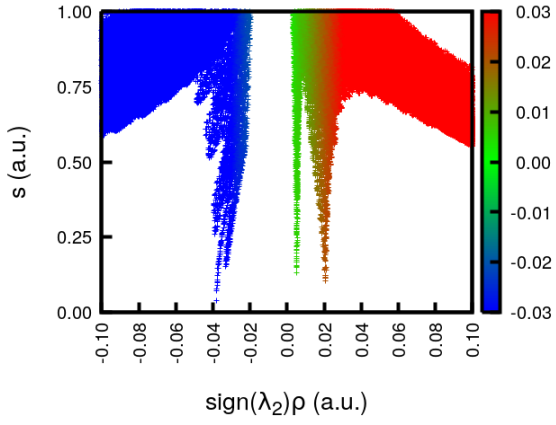
After the decomposition of the density into M monomer contributions, it is necessary to identify the particular regions Ω_{NCI} that are associated with intermolecular interactions. To



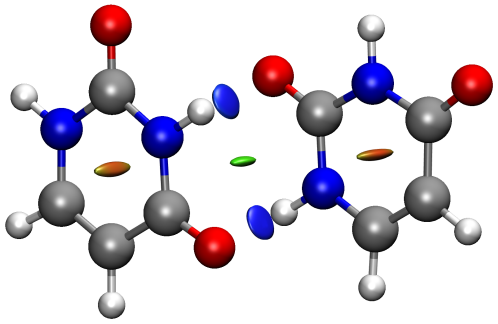
(a) 2D promolecular



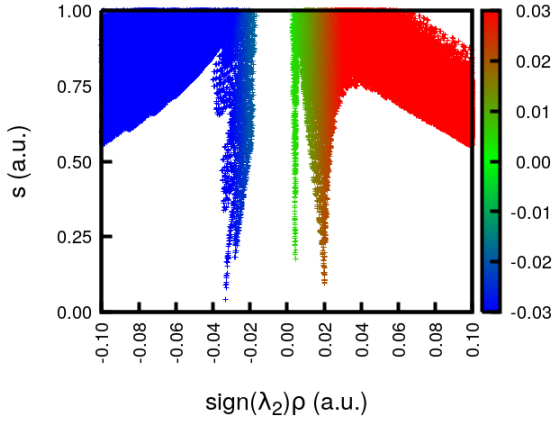
(b) 3D promolecular



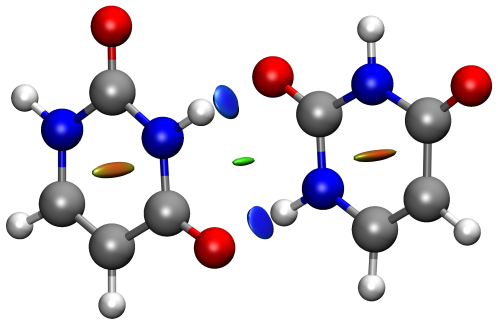
(c) 2D B3LYP



(d) 3D B3LYP



(e) 2D ELMO



(f) 3D ELMO

Figure 9.3: 2D and 3D NCI plots of the uracil dimer obtained with different underlying electron distributions: (a) and (b) promolecular density, (c) and (d) electron density resulting from B3LYP/6-311G(d,p) calculations, (e) and (f) electron density obtained from transferred ELMOs using the basis set 6-311G(d,p). In the 3D plots, the isosurfaces of the RDG correspond to an isovalue of 0.3 a.u. for the promolecular density case and to an isovalue of 0.5 a.u. for the cases of the ELMO and DFT densities. Note that a higher (or lower) isovalue only leads too larger (or smaller) isosurfaces, but the overall shape remains.

this purpose, the following two conditions are introduced:

$$\boldsymbol{r} \in \Omega_{NCI} \begin{cases} s(\boldsymbol{r}) < s^{ref}(\boldsymbol{r}) \\ \rho_{total}(\boldsymbol{r}) < \gamma^{ref}\rho_i(\boldsymbol{r}) \text{ for all } i \end{cases} \tag{9.3}$$

In other words, one has to consider only those points of the real space for which the RDG is below a certain threshold, and for which more than only one monomer contributes to the total density. For $s^{ref}(\boldsymbol{r})$, a value of 1.0 a.u. is typically used because the corresponding isosurfaces generally encompass all relevant interaction regions, which are covalent and non-covalent. Then the second restriction is introduced to keep only those points that are associated with intermolecular interactions, while regions of intramolecular interactions are not considered. The threshold γ^{ref} is an empirical value, which was initially set to 0.95 for studying the intermolecular interactions in dimers ($M = 2$ in Equation (9.2)).[375,414]

Using the definition of the intermolecular interaction region Ω_{NCI}, integrals of powers of the electron density can be computed:

$$I_n = \int_{\Omega_{NCI}} \rho^n(\boldsymbol{r}) d\boldsymbol{r} \tag{9.4}$$

To define the best combination of parameters (namely, the electron density exponent n, the density threshold γ^{ref} and the RDG threshold $s^{ref}(\boldsymbol{r})$), the correlation between the NCI integral values and interaction energies has been evaluated for the S66 database.[414] This dataset[396] contains the optimized geometries and interaction energies obtained at CCSD(T)/CBS level (exploiting the supermolecular approach with counterpoise correction) for 66 small representative dimers. The final set of parameters to compute the NCI integral values was the one providing the best correlation with the benchmark interaction energies. Boto *et al.* reported that the best correlation has been obtained for an exponent of $n = 2.5$, resulting in a correlation coefficient of 0.94.[414] However, these results were only obtained for promolecular densities. Since this approximation is very crude, other options for calculating the NCI integral values and the underlying electron densities should be considered. Therefore, we have evaluated the possibility to quantify interaction strengths using the NCI-ELMO approach. The corresponding study has not been published yet, but the results are nevertheless presented in Section 9.3.

9.2 Qualitative analysis

In this section, the results obtained with the NCI-ELMO, NCI-DFT and promolecular-NCI techniques will be compared on a qualitative basis for a variety of non-covalent interactions that typically occur in biomolecules. As mentioned above, the results discussed in this section were already published in reference [320].

9.2.1 Computational details

In the following, an overview about the studied systems and types of interactions will be given alongside with a description of the procedure to pre-process the corresponding PDB files, to

compute the densities and to perform the NCI analyses.

9.2.1.1 Selected polypeptides and proteins

The different NCI computations were performed on the following systems and types of interactions.

- Polypeptides (benchmark systems):
  - a strong hydrogen bond,
  - a weak hydrogen bond, and
  - interactions with a metal center.
- Proteins:
  - a strong hydrogen bond,
  - a C−H$\cdots\pi$ interaction,
  - a cation$\cdots\pi$ interaction,
  - an anion$\cdots\pi$ interaction,
  - lone pair (n−π^*) interactions, and
  - interactions with a metal center.

9.2.1.2 Pre-processing of the structures

The structures of the analyzed proteins and polypeptides were mostly taken from the PDB.[219] When necessary, the following modifications were made to the PDB files: (i) only one of the disordered conformations of the polypeptide or protein was chosen; (ii) the correct protonation states were assigned according to the pH value of crystallization or nuclear magnetic resonance (NMR) experiment; and (iii) missing hydrogen atoms were added using the *tleap* software in *AMBER*.[148]

9.2.1.3 Density computation and subsequent NCI analysis

An initial validation of the NCI-ELMO technique was performed on polypeptides because they represent a good compromise. In fact, they include the same types of interactions found in proteins, but, due to their medium sizes, fully quantum mechanical calculations are still feasible. In particular, the non-covalent interactions were analyzed based on densities obtained from (i) transfers of ELMOs from the ELMO libraries (compare Section 1.3); (ii) DFT calculations; and (iii) the promolecular approximation. Transfers of ELMOs and DFT calculations were performed exploiting all five basis sets currently available in the *ELMOdb* program (6-31G, 6-311G, 6-31G(d,p), 6-311G(d,p), and cc-pVDZ). DFT calculations were performed using the B3LYP functional for the polypeptides containing only amino acids and water, or with the B3PW91 functional[430,431] for the polypeptide with a metal center. However, since DFT calculations are impractical for proteins, for these larger systems the NCI-ELMO results were compared only to those obtained with the promolecular-NCI approach.

The ELMO transfers were performed with the *ELMOdb* software,[105] and the DFT calculations with *Gaussian09*[150]. In both cases, the programs provided a wavefunction file (in wfx format) that was passed to the *NCIPLOT* software[426] to compute the density and

the reduced density gradient. The computation of the promolecular densities is implemented directly in the *NCIPLOT* program.

To study the interactions with the metal centers, it was necessary to compute tailor-made ELMOs for each Zn^{2+} ion on suitable model molecules that mimic the chemical environments of the metal in the investigated systems. These calculations were performed with our in-house modified version[103] of *GAMESS-UK*[118] that implements the Stoll technique for the computation of ELMOs (for more details about the computation of ELMOs, see Section 1.3.2). For the construction the promolecular densities of the polypeptide and protein with the metal centers, the spherical atomic density was specifically computed for the charged Zn^{2+} ion. This is contrary to the usual procedure of calculating promolecular densities, which are normally obtained from the spherical densities (of the neutral atoms) that are stored in the *NCIPLOT* program.

9.2.2 Validation on polypeptides

To validate the new NCI-ELMO technique, the NCI results obtained for promolecular, DFT and ELMO densities were compared for three different polypeptides, which were chosen as test cases because they include common non-covalent interactions that are also found in proteins.

As a first test case, we studied a strong hydrogen bond in Leu-enkephalin. The structure of this polypeptide was obtained from a refinement of high-resolution X-ray data.[314] A hydrogen bond is formed between residues Tyr2 and Phe5, and corresponds to the typical disk-shaped RDG isosurface shown in Figure 9.4A. The actual comparison between the different densities was performed through an analysis of the 2D NCI plots. The plots for the promolecular, B3LYP and ELMO electron densities are simultaneously shown in Figure 9.4B for basis set 6-31G and in Figure 9.4C for the cc-pVDZ set of basis functions. In all the considered cases, a single peak is observed in the negative region of the "signed" electron density. This is typical for strong hydrogen bonds.[429]

Nevertheless, in panels B and C of Figure 9.4, the peaks for the promolecular density are shifted to more negative values of the "signed" electron density compared to the B3LYP ones, which are clearly more similar to the ELMO peaks. Analogous results are obtained for the other three basis sets, as can be seen in Figure C.1 in the Appendix. A direct comparison of the different basis sets is shown in Figures 9.5A and B for NCI-B3LYP and NCI-ELMO, respectively. In both cases, the peak corresponding to the strong hydrogen bond is shifted to more negative values when the basis set becomes larger.

As a second benchmark test, we focused on a weak hydrogen bond in lactoferrampin, an antimicrobial peptide characterized by antibacterial and candidacidal activities.[432] The structure of the polypeptide was measured by NMR and only the first geometry in the PDB file 2MD3 was considered in this study. The weak hydrogen bond is formed between residues Phe2 and Gly3 and is depicted through a typical green RDG isosurface shown in Figure 9.6A. The corresponding 2D plots are shown in Figures 9.6B and 9.6C for the promolecular, B3LYP and ELMO levels. For the last two methods, Figure 9.6B shows the results obtained with the basis set 6-31G, while in Figure 9.6C the plots obtained for the cc-pVDZ set of basis functions are depicted.

For the weak hydrogen bond, all the different levels provide two symmetrical peaks with

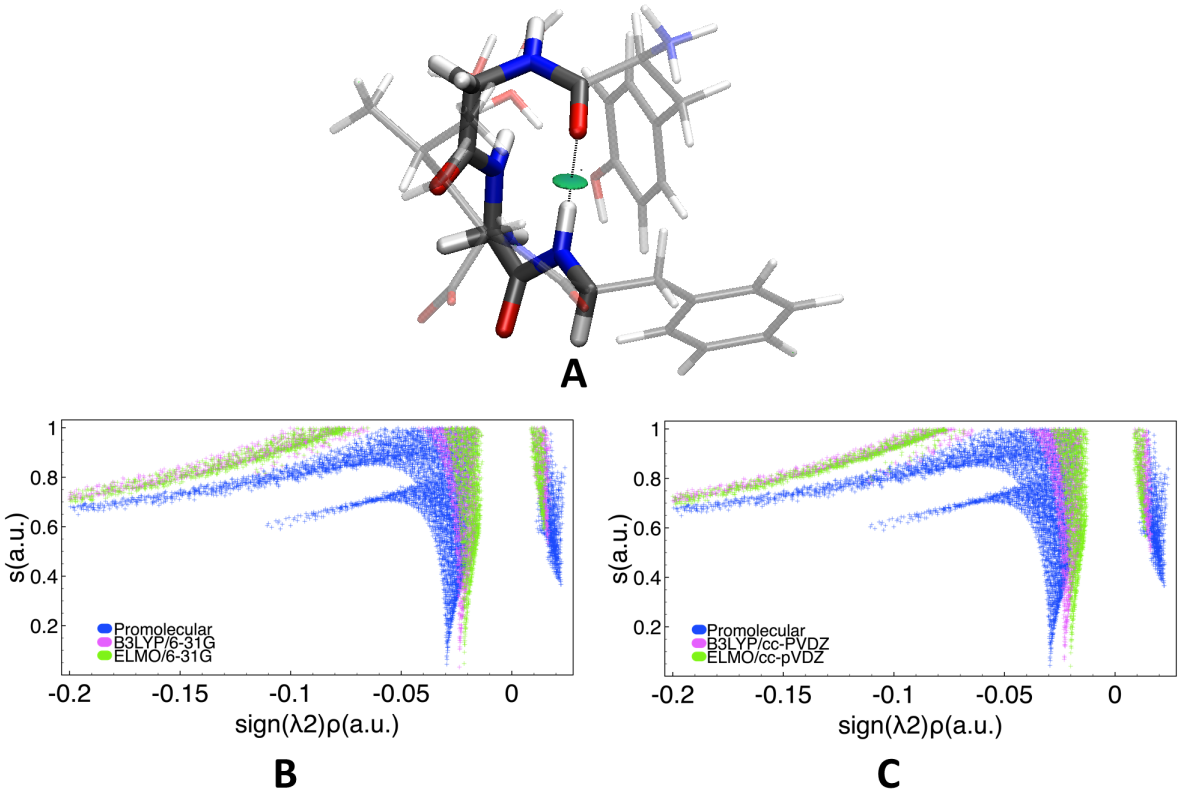


Figure 9.4: Strong hydrogen bond between Tyr2 and Phe5 in Leu-enkephalin:[314] (A) RDG isosurface ($s = 0.6$ a.u., color scale: -0.03 a.u. $< sign(\lambda_2)\,\rho <$ 0.03 a.u.) obtained at NCI-ELMO level; 2D NCI plots obtained at promolecular-NCI, NCI-B3LYP and NCI-ELMO levels for basis sets (B) 6-31G and (C) cc-pVDZ. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

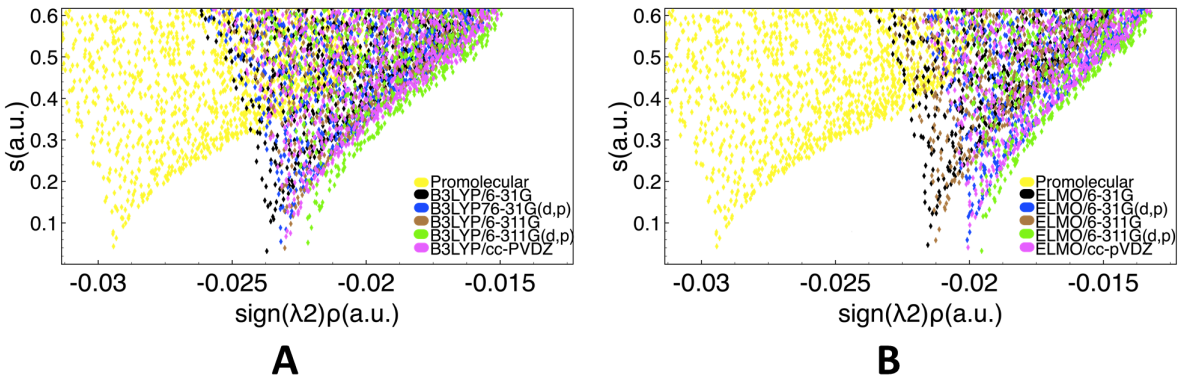


Figure 9.5: Strong hydrogen bond between Tyr2 and Phe5 in Leu-enkephalin[314]: 2D NCI plots obtained at (A) NCI-B3LYP and (B) NCI-ELMO levels with all the considered basis sets. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

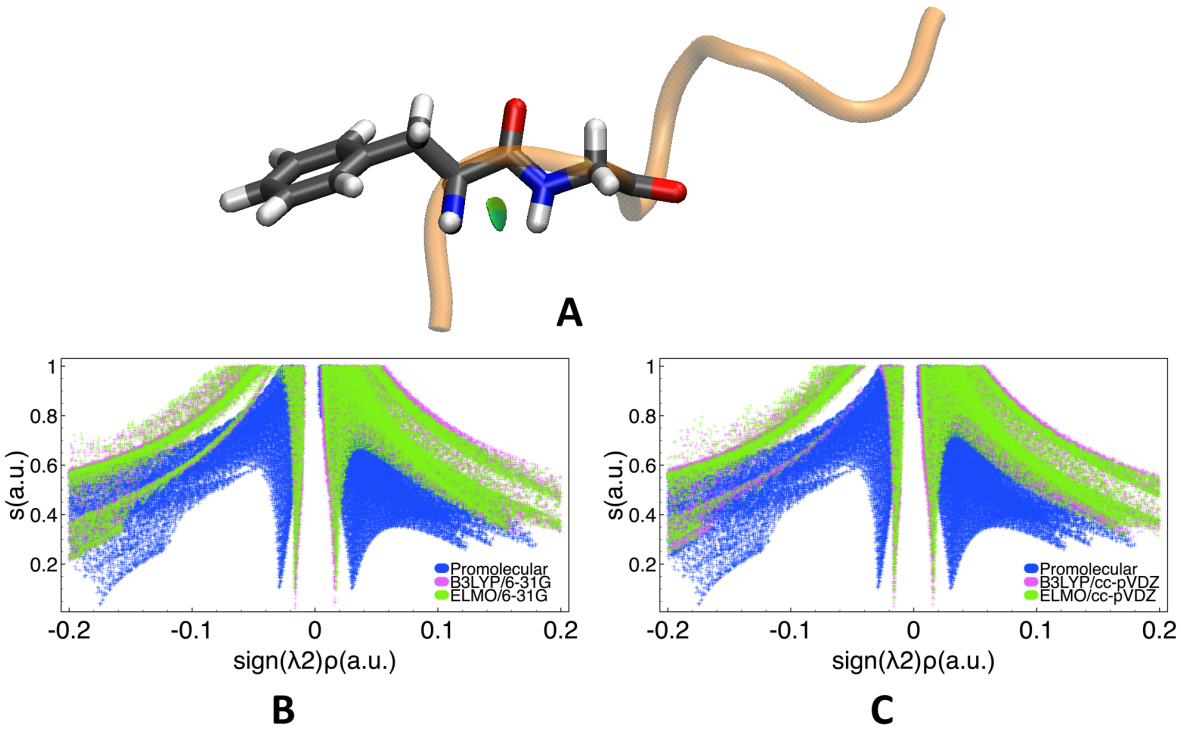


Figure 9.6: Weak hydrogen bond between Phe2 and Gly3 in lactoferrampin (PDB code: 2MD3): (A) RDG isosurface (s=0.6 a.u., color scale: -0.03 a.u. $< sign(\lambda_2)\,\rho <$ 0.03 a.u.) obtained at NCI-ELMO level; 2D NCI plots obtained at promolecular-NCI, NCI-B3LYP and NCI-ELMO levels for basis sets (B) 6-31G and (C) cc-pVDZ. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

low absolute density values, which is typical for weak interactions.[429] However, the peaks that are associated with B3LYP and ELMO densities are practically indistinguishable, while the ones for the promolecular density are clearly shifted to higher absolute values. Moreover, the B3LYP and ELMO peaks are much narrower than the promolecular ones. Identical results were also obtained for the other three basis sets (compare Figure C.2 in the Appendix). In Figure 9.7, the 2D NCI plots are compared directly for the different basis sets used in the computation of the B3LYP and ELMO densities. For both methods, the peaks overlap completely, and all the basis sets yield similar results.

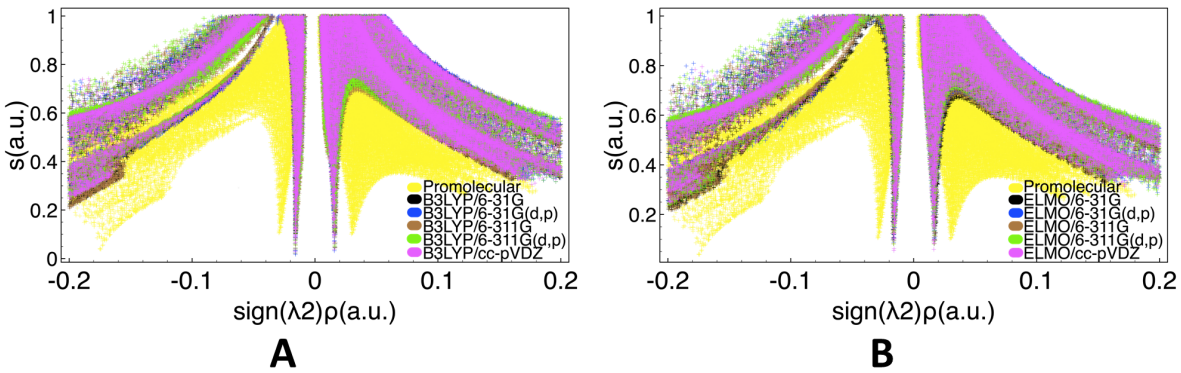


Figure 9.7: Weak hydrogen bond between Phe2 and Gly3 in lactoferrampin (PDB code: 2MD3): 2D NCI plots obtained at (A) NCI-B3LYP and (B) NCI-ELMO levels with all the considered basis sets. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

For the third and last benchmark test, interactions with a metal center were evaluated. In particular, we considered a HIV Zinc fingerlike domain (PDB code: 2ZNF), where the metal center (Zn^{2+}) is tetra-coordinated by one histidine and three cysteine residues. These interactions were the most challenging of the three benchmark tests because they are the strongest and involve a charged metal center. For their study, the first NMR molecular geometry in the PDB file 2ZNF[433] was chosen. On this structure, DFT calculations were performed with the B3PW91 functional. The ELMO density was constructed by transferring ELMOs from the database for the amino acids and tailor-made ELMOs for the Zn^{2+} ion. For the promolecular density, a spherical atomic density was specifically computed for the Zn^{2+} ion (as already mentioned in Section 9.2.1).

For each of the four coordinating residues, a blue disk-shaped RDG surface highlights the interaction with the metal center (see Figure 9.8). For a better comparison between the

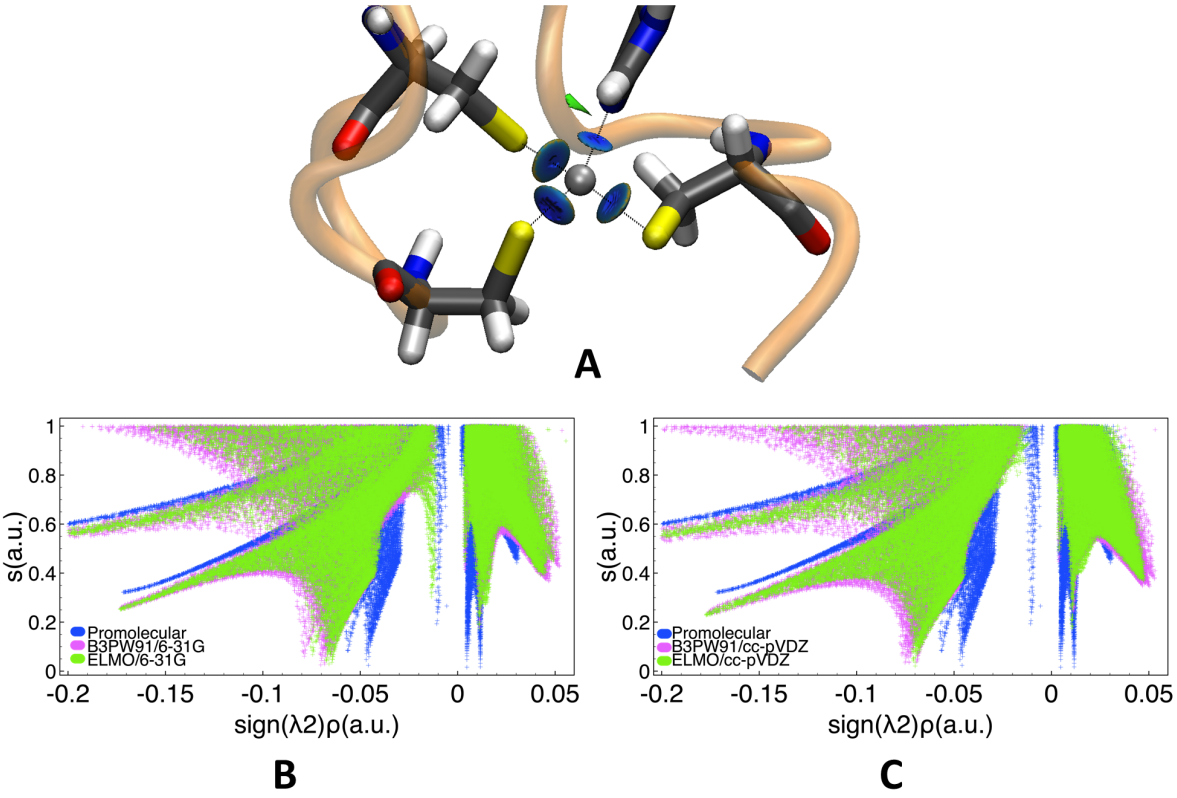


Figure 9.8: Interactions between the Zn^{2+} ion and the coordinating residues (Cys3, Cys6, Cys16 and His11) in a HIV Zinc fingerlike domain (PDB code: 2ZNF): (A) RDG isosurfaces (s=0.6 a.u., color scale: -0.03 a.u. $< sign(\lambda_2)\,\rho <$ 0.03 a.u.) obtained at NCI-ELMO level; 2D NCI plots obtained at promolecular-NCI, NCI-B3PW91 and NCI-ELMO levels for basis sets (B) 6-31G and (C) cc-pVDZ. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

different densities, the 2D NCI plots are shown for all levels of theory in Figure 9.8B for the 6-31G basis set and in Figure 9.8C for the cc-pVDZ set of basis functions. For both basis sets it is evident that the promolecular approximation is not able to provide the same spikes obtained through the B3PW91 reference calculations. In particular, in the negative "signed" electron density region, the promolecular peaks are closer to zero than the B3PW91 spikes. In contrast, a much better agreement is obtained between the ELMO and B3PW91 levels, although the ELMO peaks occur at slightly smaller absolute values of the "signed" electron

density. Analogous results were obtained with the other three basis sets, for which the 2D NCI plots are shown in Figure C.3 in the Appendix.

In addition to the above-discussed NCI plots, the computational cost for each NCI analysis is reported in Table 9.1. For all the different polypeptides and basis sets, the DFT calculations are clearly the most time consuming way to obtain the densities. Depending on the size of the system and the basis set, the NCI-DFT analyses may take between 30 minutes or more than two days. In contrast, the promolecular approach is clearly the fastest and is able to provide the results within seconds. However, for all the investigated polypeptides and basis sets, also the NCI-ELMO method took only few seconds, or, for the largest system, up to 13 minutes. We believe that this computational cost is still completely acceptable since the NCI-ELMO method outperforms the promolecular approximation and gives results that are very close to those obtained with DFT calculations, independently of the adopted basis set.

Table 9.1: Global CPU times (in the format hh:mm:ss.000) for the NCI calculations on the polypeptides. For the NCI-ELMO and NCI-DFT cases, the global CPU time is the sum of the time required to compute the wavefunction with *ELMOdb* or *Gaussian09* plus the time for the NCI analysis taken by the *NCIPLOT* program. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

	Leu-enkephalin	Lactoferrampin	HIV Zinc fingerlike domain
Number of atoms	86	145	275
Number of electrons	326	522	1098
Promolecular-NCI	00:00:00.214	00:00:00.977	00:00:01.595
NCI-ELMO/6-31G	00:00:04.316	00:01:09.809	00:08:16.113
NCI-ELMO/6-311G	00:00:05.758	00:01:16.750	00:10:00.541
NCI-ELMO/6-31G(d,p)	00:00:05.980	00:01:43.015	00:10:41.436
NCI-ELMO/6-311G(d,p)	00:00:07.321	00:01:49.644	00:11:58.663
NCI-ELMO/cc-pVDZ	00:00:06.964	00:01:52.870	00:12:40.584
NCI-DFT/6-31G	00:31:31.273	02:29:19.071	04:46:49.380
NCI-DFT/6-311G	00:37:43.427	04:03:45.798	17:18:36.688
NCI-DFT/6-31G(d,p)	00:58:55.350	08:20:50.301	38:56:23.140
NCI-DFT/6-311G(d,p)	02:45:32.024	17:39:25.391	47:31:56.169
NCI-DFT/cc-pVDZ	01:55:34.030	20:53:03.176	51:36:05.193

9.2.3 Application to proteins

Encouraged by the promising results for the polypeptides, we decided to apply the new NCI-ELMO method to a range of different non-covalent interactions in proteins, for which the results will be discussed in the following subsections.

9.2.3.1 Hydrogen bonds

As a first example for non-covalent interactions in proteins, we analyzed two different hydrogen bonds using the NCI-ELMO and NCI-promolecular approaches. Despite the concept of hydrogen bonding has been known for a long time,[351] the exact definition varied over time, and has been renewed in 2011 by the International Union of Pure and Applied Chemistry.[350] This new definition does not specify the element types of hydrogen bond donors and ac-

ceptors and therefore includes also less conventional hydrogen bonds.[351] To take this into account, in our study[320] we chose to investigate a conventional O−H···O interaction and a less conventional C−H···π interaction.

As an example for a conventional hydrogen bond, we have analzed the O−H···O interaction between residues Asp75 and Asp87 in the structure of the D192N mutant of Rhamnogalacturonan acetylesterase[434] (PDB code 3C1U). The presence of this interaction was previously indicated in the corresponding ^{1}H-NMR spectrum.[434] To characterize this interaction, promolecular-NCI and NCI-ELMO analyses were performed, the latter exploiting all the basis sets available in the libraries. The hydrogen bond is revealed in the NCI analysis by the disk-shaped RDG isosurface in the 3D NCI plot in Figure 9.9A and by the peak in the negative region of the "signed" electron density of the 2D NCI plot in Figure 9.9B. As previ-

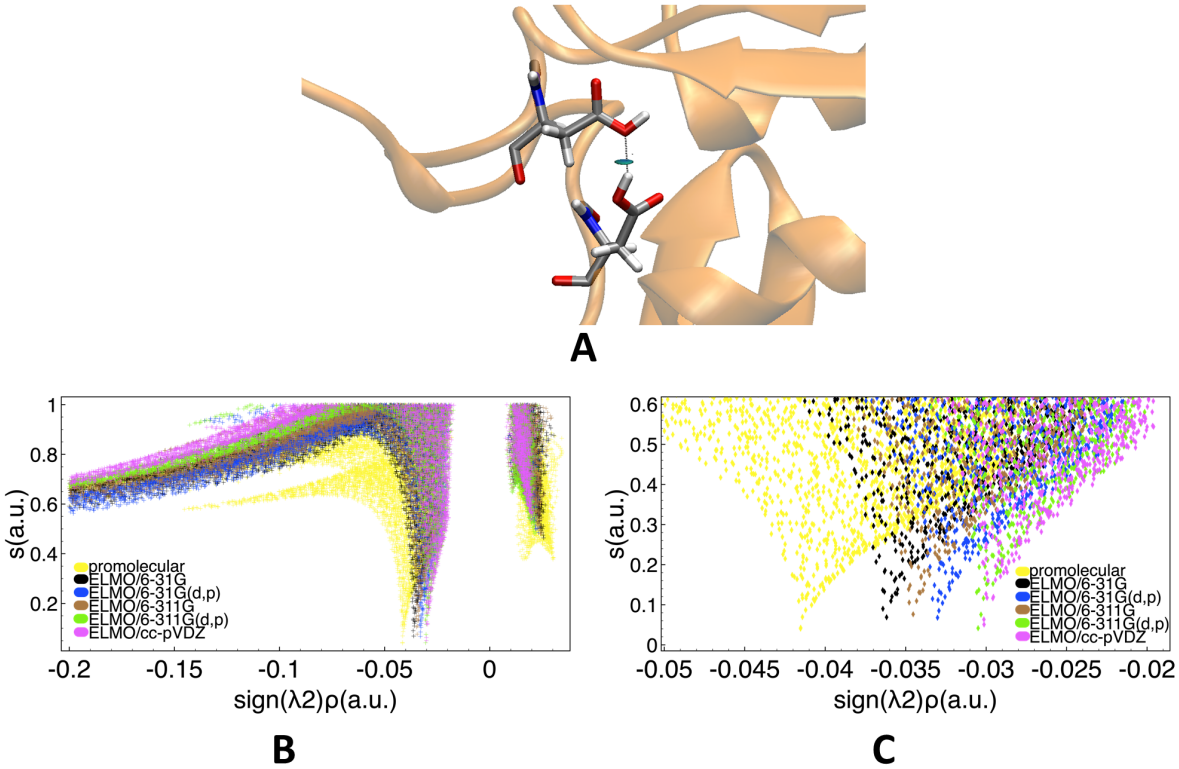


Figure 9.9: Strong hydrogen bond between residues Asp75 and Asp87 in the D192N mutant of Rhamnogalacturonan acetylesterase (PDB code: 3C1IU): (A) RDG isosurface (s=0.6 a.u., color scale: -0.03 a.u. $< sign(\lambda_2)\,\rho <$ 0.03 a.u.) obtained at NCI-ELMO level; (B) 2D NCI plots obtained at promolecular-NCI and NCI-ELMO levels for all the considered basis sets; (C) Zoomed 2D NCI plots obtained at promolecular-NCI and NCI-ELMO levels for all the considered basis sets. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

ously mentioned in the discussion of the strong hydrogen bond in Leu-enkephalin, this peak is typical for strong hydrogen bonds. In analogy to what has already been observed for Leu-enkephalin (compare Figures 9.4B and 9.4C), also in this case the promolecular approximation corresponds to a peak that is shifted toward more negative values of the "signed" electron density. Concerning the different basis sets used in the NCI-ELMO computations, the peaks are shifted toward less negative values with increasing size and quality of the chosen basis set, in the order 6-31G, 6-311G, 6-31G(d,p), 6-311G(d,p) and cc-pVDZ. Except for the cc-pVDZ

basis set, the same trend was also observed for Leu-enkephalin (compare Figure 9.5). These observations show the high similarity between the two strong hydrogen bonds in the studied polypeptide and protein structures.

We have then analyzed a less conventional hydrogen bond, namely a C−H···π interaction, whose presence has been detected by NMR spectroscopy between the methyl group of Leu50 and the aromatic ring of Tyr59 in the human erythrocytic ubiquitin.[435] The NCI analysis was based on the corresponding crystallographic structure of the protein[436] (PDB code: 1UBQ). Apart from further assessing the capabilities of the new NCI-ELMO method, the goal of this analysis is to investigate to which extent the C−H···π interaction is comparable to more traditional types of hydrogen bonds.

In the 3D NCI plot (Figure 9.10A), the C−H···π interaction corresponds to a green RDG isosurface that is delocalized due to the interaction of the hydrogen atom with the whole π-electron cloud of the aromatic ring.[427] The corresponding 2D NCI plots are shown in Figure 9.10B. They are very similar to the plots obtained for the weak hydrogen bond in lactoferrampin (see Figure 9.7B). For both systems, the interaction is characterized by two symmetrical peaks around zero, which are shifted to higher absolute values for promolecular densities compared to the ELMO ones. Identical results are practically obtained for all five basis sets underlying the NCI-ELMO computations, which is again in analogy with the findings for the polypeptide lactoferrampin. In conclusion, the C−H···π interaction yields similar NCI results to a conventional weak hydrogen bond, confirming the similarity of the two interactions.

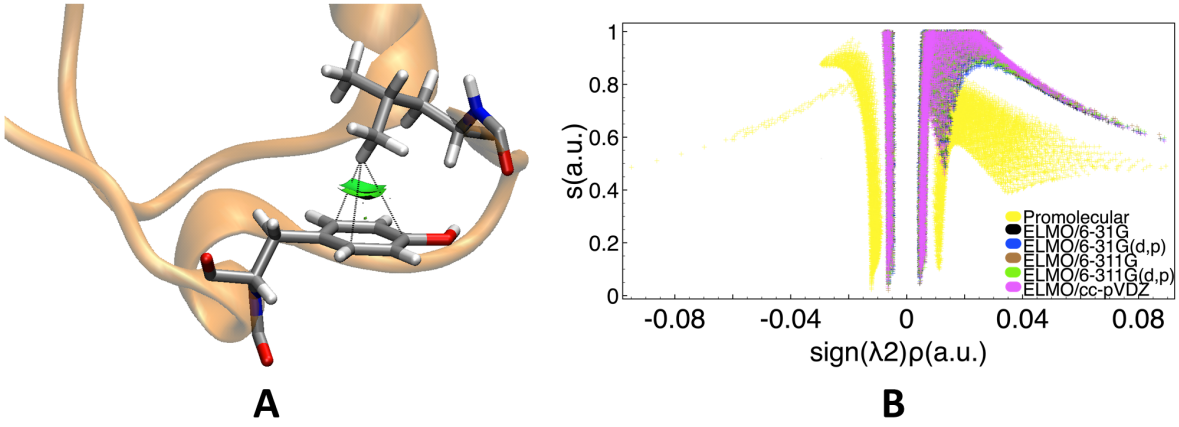


Figure 9.10: C−H···π interaction between Leu50 and the aromatic ring in Tyr59 in the human erythrocytic ubiquitin (PDB code: 1UBQ): (A) RDG isosurface (s=0.6 a.u., color scale: -0.03 a.u. $< sign(\lambda_2)\,\rho <$ 0.03 a.u.) obtained at NCI-ELMO level; (B) 2D NCI plots obtained at promolecular-NCI and NCI-ELMO levels for the considered basis sets. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

Moreover, considering the comparisons of the NCI results obtained for the three different polypeptides and the two proteins, the basis set seems to have only little influence on the outcome of the analyses. Hence, it can be concluded that the choice of method for the computation of the density is more important than the choice of the basis set. Therefore, in the study of the following proteins, only one set of basis functions (namely cc-pVDZ) will be used to compute the ELMO densities.

9.2.3.2 Cation$\cdots\pi$ and anion$\cdots\pi$ interactions

Another common class of non-covalent interactions in proteins includes cation$\cdots\pi$[356,357,437] and anion$\cdots\pi$[358,359] interactions. The former are known to play an important role in protein structures and protein-ligand interactions,[437] while the latter are a rather new type of interactions that are expected to be important for protein folding and function.[359] In proteins, cation$\cdots\pi$ interactions are formed between aromatic side chains (histidine, phenylalanine, tyrosine or tryptophan) and positively charged residues (arginine, lysine or histidine). Depending on its protonation state, histidine can act either as a cation or as π-system.[357] Most of the anion$\cdots\pi$ interactions in proteins are formed between the same aromatic side chains that are also involved in the cation$\cdots\pi$ interactions and negatively charged residues (aspartate and glutamate).[359]

To further evaluate the capabilities of the NCI-ELMO technique, it has been applied to study a protein where one lysine residue is involved in four cation$\cdots\pi$ interactions with four aromatic side chains. In particular, this interaction occurs in the X-ray structure of a complex formed by glucoamylase and D-*gluco*-dihydroacarbose (PDB code: 1GAI)[438] between the positively charged NH_3^+ group of Lys108 and the aromatic side chains of Trp52, Trp120, Tyr50 and Tyr116. All four interactions were identified by the NCI-ELMO technique and correspond to the delocalized RDG isosurfaces of green color shown in Figure 9.11A.

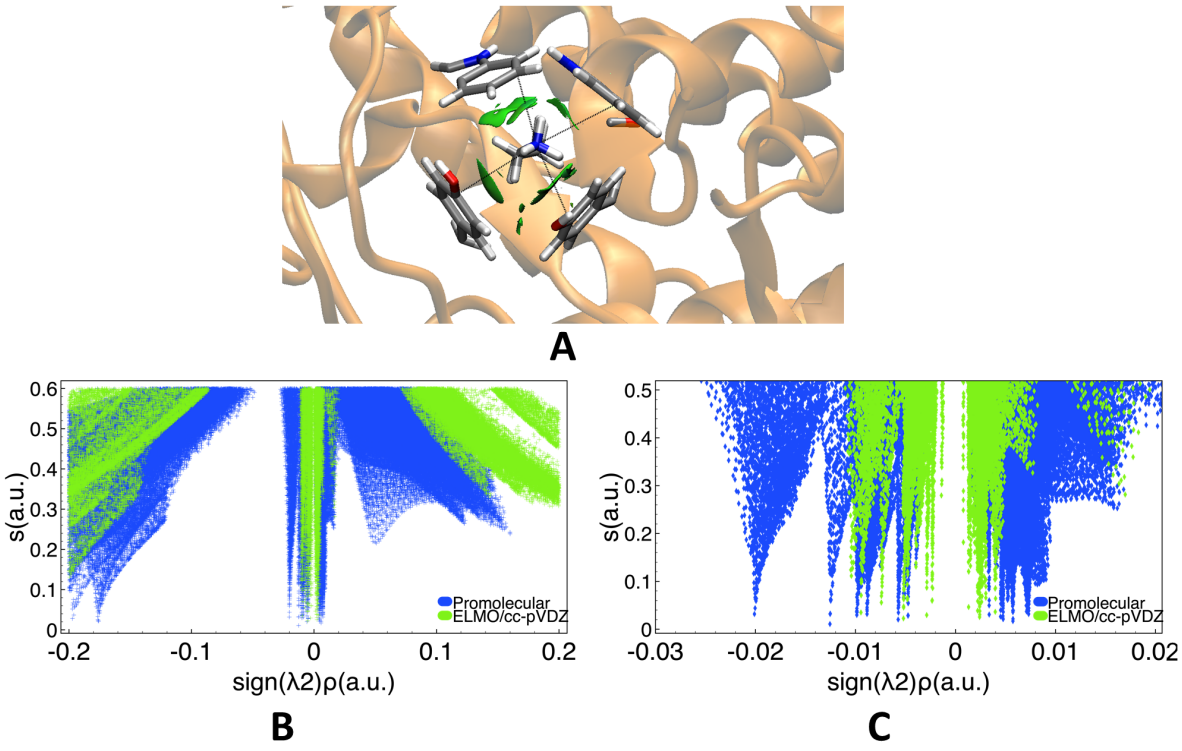


Figure 9.11: Cation$\cdots\pi$ interaction between the positively charged ammonium group of Lys108 with the aromatic sidechains of Trp52, Trp120, Tyr50 and Tyr116 in the complex of glucoamylase with D-*gluco*-dihydroacarbose (PDB code: 1GAI): (A) RDG isosurfaces (s=0.6 a.u., color scale: -0.03 a.u. $< sign(\lambda_2)\,\rho <$ 0.03 a.u.) obtained at NCI-ELMO level; (B) 2D NCI plots obtained at promolecular-NCI and NCI-ELMO/cc-pVDZ levels; (C) Zoomed 2D NCI plots obtained at promolecular-NCI and NCI-ELMO/cc-pVDZ levels. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

In addition to the NCI-ELMO calculation (with the cc-pVDZ basis set), a promolecular-NCI computation was also performed. The two levels are compared directly in the 2D NCI plots, which are reported in Figures 9.11B (general plot) and 9.11C (zoomed plot). Due to the presence of the four simultaneously occurring interactions, several different peaks can be identified in the 2D plots. Independently of the underlying density, sharp peaks are observed at low absolute values of the "signed" electron density. However, as also seen for the hydrogen bonds, the promolecular peaks are shifted to higher absolute density values compared to the NCI-ELMO/cc-pVDZ ones.

Furthermore, we also analyzed the interaction of an anion with a π-system. The promolecular and ELMO/cc-pVDZ densities were computed for the crystal structure of the antigene-binding fragment of the catalytic antibody 15A9 in complex with phosphopyridoxyl-*L*-alanine[439] (PDB code: 1WC7). In this system, the anion$\cdots\pi$ interaction occurs between Glu58 and Tyr94 and has been identified by Lucas *et al.*[359] based on geometrical criteria. Also the NCI method reveals the interaction in form of a deleocalized, green RDG isosurface (Figure 9.12A). Figure 9.12B shows the corresponding 2D plots resulting from the NCI-ELMO and promolecular-NCI analyses, which have the typical shape of weak non-covalent interactions, with the promolecular peaks being again shifted to higher absolute values of the "signed" electron density compared to the ELMO ones. Additionally, also in this case, the ELMO peaks are much narrower than the promolecular ones.

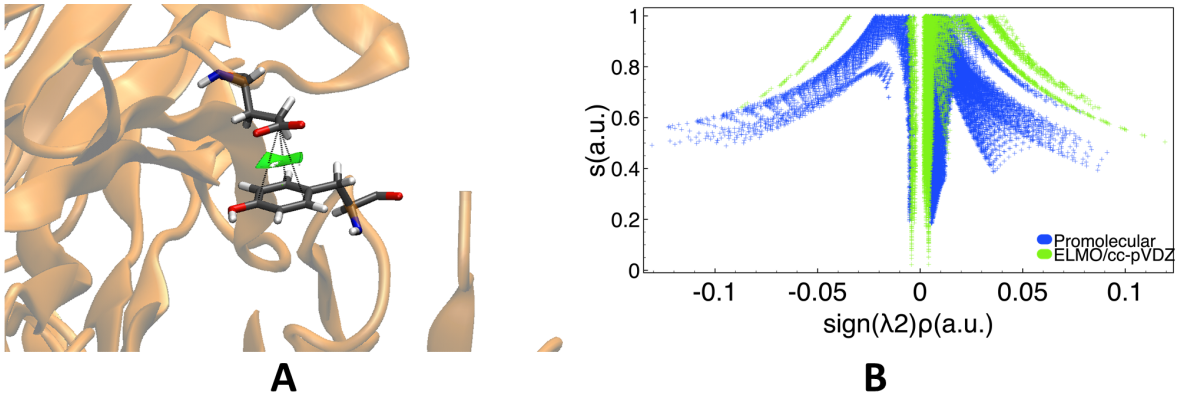


Figure 9.12: Anion$\cdots\pi$ interaction between Glu58 and Tyr94 in the antigene-binding fragment of the catalytic antibody 15A9 in complex with phosphopyridoxyl-*L*-alanine (PDB code: 1WC7): (A) RDG isosurface (s=0.6 a.u., color scale: -0.03 a.u. $< sign(\lambda_2)\,\rho <$ 0.03 a.u.) obtained at NCI-ELMO level; (B) 2D NCI plots obtained at promolecular-NCI and NCI-ELMO/cc-pVDZ levels. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

9.2.3.3 Lone pair (n-π*) interactions

The n-π* interaction is a relatively new and emerging type of non-covalent interaction where a lone pair (n) of a carbonyl oxygen overlaps with the antibonding orbital (π*) of another carbonyl group.[440] In proteins, half of the non-hydrogen atoms in the backbone belong to carbonyl groups[441] and an analysis of high-resolution crystal structures in the PDB indicated that, based on geometrical criteria, n-π* interactions could be present in all of the examined protein structures and occur in approximately 34% of all residues.[379]

Futhermore, since the lone pair atoms of a carbonyl oxygen atom can be simultaneously involved in hydrogen bonds and n-π^* interactions, the two interactions influence each other. To study this interplay, Bartlett *et al.* focused on the carbonyl oxygen in the side chain of asparagine. This atom can act as an acceptor of a hydrogen bond and, at the same time, also as a donor of a n-π^* interaction that is formed with the backbone carbonyl carbon of the same asparagine residue.[380] In their study, Bartlett *et al.* concluded that n-π^* interactions, together with hydrogen bonds, make an important contribution to the stability of protein structures.[380]

In particular, they discovered that weak hydrogen bonds are accompanied by strong n-π^* interactions. Furthermore, the presence of n-π^* interactions seems to depend on the kind of hydrogen bond. In fact, the carbonyl oxygen atoms in the side chains of asparagine residues were found to establish n-π^* interactions mostly in combination with the following two different kinds of hydrogen bonds: (i) local ones, which occur between almost neighboring residues (namely between residue i and residue $i + 2$); and (ii) non-local hydrogen bonds, for which the two hydrogen bonded residues are separated by at least five intermeditate residues.

Accordingly, we have decided to study both kinds of interaction motifs, with the goal of investigating (i) whether the NCI-ELMO (and promolecular-NCI) technique is able to reveal the presence of the n-π^* interactions; and (ii) whether any differences between the two kinds of motifs could be observed also from the NCI analysis. To accomplish these tasks, the promolecular-NCI and NCI-ELMO (cc-pVDZ basis set) techniques were applied to two different protein crystal structures. In human carbonic anhydrase II[442] (PDB code: 3KS3), a local hydrogen bond is formed between residues Asn61 and Gly63 and is accompanied by an n-π^* interaction within residue Asn61. The corresponding non-local interaction is present in leucyl/phenylalanyl-tRNA-protein transferase[443] (PDB code: 2CXA). In particular, the non-local hydrogen bond is formed between residues Asn24 and Asp62, while the n-π^* interaction takes place in Asn24. The interactions in these two systems were also investigated by Bartlett *et al.* by applying natural bond orbital (NBO) analyses.[380]

The NCI analysis of both structures identified two interactions for each of the carbonyl oxygen atoms, as can be seen in the NCI-ELMO 3D plots reported in Figure 9.13. In particular, both kinds of hydrogen bonds appear as two light-blue disk-shaped RDG isosurfaces, while the n-π^* interactions correspond to the green isosurfaces. Interestingly, the RDG isosurface for the n-π^* interaction in 3KS3 (the one accompanying the local hydrogen bond) is much larger than the isosurface corresponding to the n-π^* interaction in protein 2CXA.

In Figures 9.14A and 9.14C, the 2D plots corresponding to the local and non-local hydrogen bonds are shown, respectively. Both plots are very similar to the ones obtained for other conventional hydrogen bonds (compare Figures 9.4C and 9.9B) and also the differences between the promolecular-NCI and NCI-ELMO techniques are identical to those observed for the other systems. In fact, also in the cases of the local and non-local hydrogen bonds, the promolecular peaks are shifted to lower values of the "signed" electron density compared to the NCI-ELMO ones.

Furthermore, the 2D NCI plots for the corresponding n-π^* interactions are shown in Figures 9.14B and 9.14D for the protein structures 3KS3 and 2CXA, respectively. Both interactions are clearly identified as weak, with the typical shape already observed for the

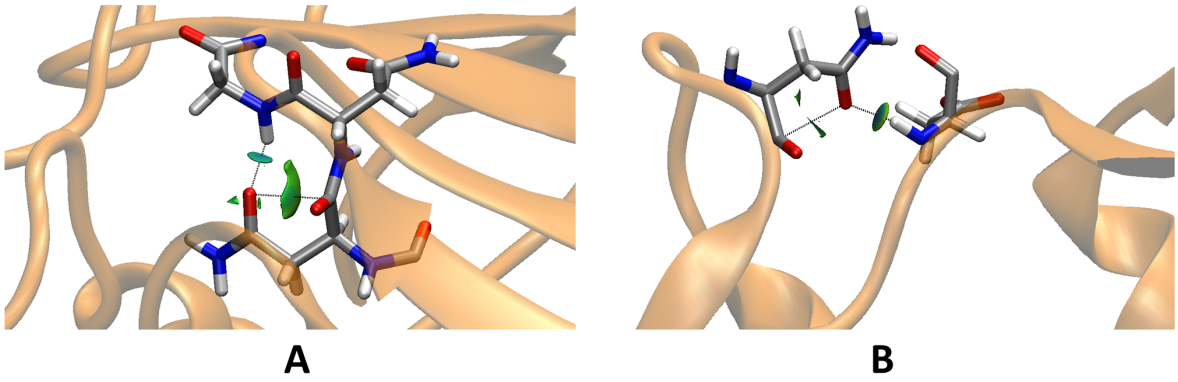


Figure 9.13: RDG isosurfaces (s=0.6 a.u., color scale: -0.03 a.u. $< sign(\lambda_2)\,\rho <$ 0.03 a.u.) obtained at NCI-ELMO level: (A) local hydrogen bond between residues Asn61 and Gly63 and n-π^* interaction between the oxygen atom (OD1) in the side chain of Asn61 and the carbon atom (C) of the carboxylic group in the backbone of Asn61 in the human carbonic anhydrase II protein (PDB code: 3KS3); (B) non-local hydrogen bond between Asn24 and Asp62 and n-π^* interaction between the oxygen atom (OD1) in the side chain of Asn24 and the carbon atom (C) of the carboxylic group in the backbone of Asn24 in the leucyl/phenylalanyl-tRNA-protein transferase (PDB code: 2CXA). Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

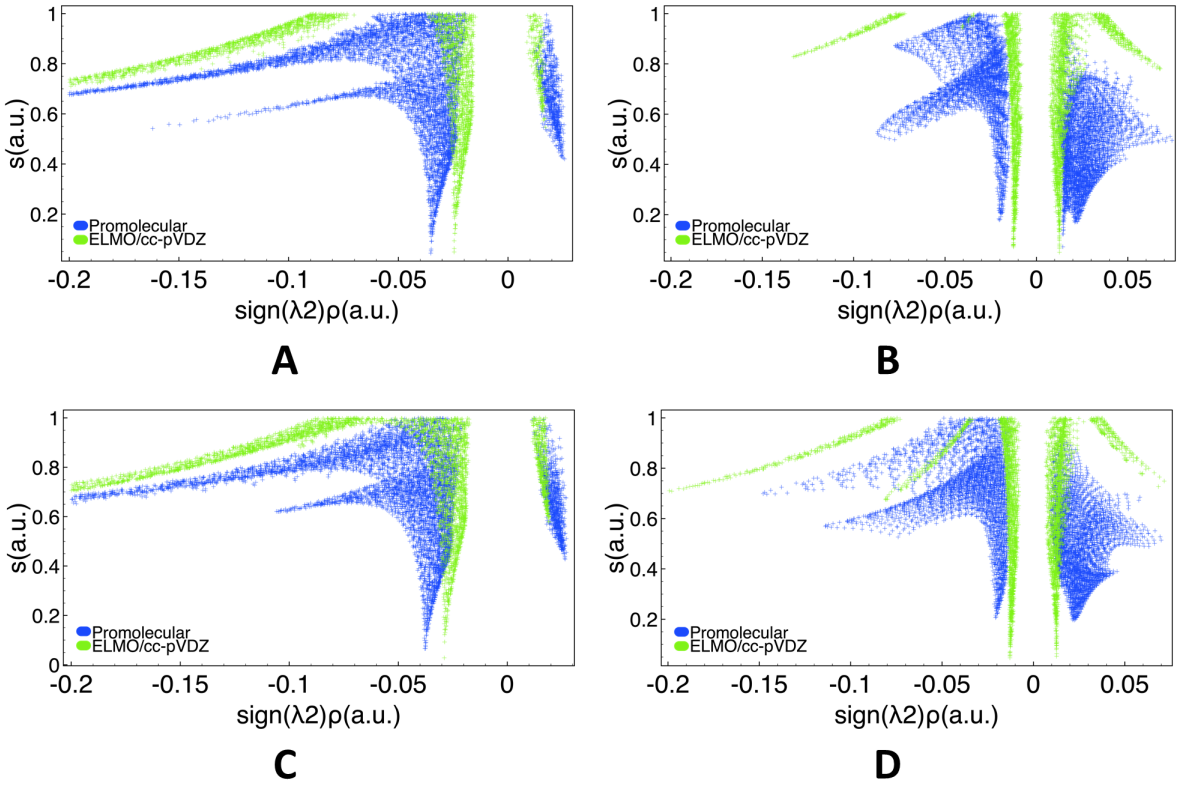


Figure 9.14: 2D NCI plots obtained at promolecular-NCI and NCI-ELMO/cc-pVDZ levels: (A) local hydrogen bond between residues Asn61 and Gly63 and (B) n-π^* interaction between the oxygen atom (OD1) in the side chain of Asn61 and the carbon atom (C) of the carboxylic group in the backbone of Asn61 in the human carbonic anhydrase II protein (PDB code: 3KS3); (C) non-local hydrogen bond between Asn24 and Asp62 and (D) n-π^* interaction between the oxygen atom (OD1) in the side chain of Asn24 and the carbon atom (C) of the carboxylic group in the backbone of Asn24 in the leucyl/phenylalanyl-tRNA-protein transferase (PDB code: 2CXA). Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

other weak non-covalent interactions in this study (compare Figures 9.6C and 9.10B). For all these weak interactions, including the n-π^* interactions, the same trends can be observed, with the promolecular peaks being broader, less well defined and shifted to higher absolute values compared to the NCI-ELMO ones.

Concerning the n-π^* interactions, it can be concluded that both the promolecular-NCI and the NCI-ELMO techniques are able to identify the simultaneous presence of the hydrogen bonds and n-π^* interactions. According to the 2D NCI plots, both kinds of hydrogen bonds show the typical "fingerprint" of strong interactions, while the n-π^* interactions are clearly identified as weak interactions. However, differences between the local and non-local cases can only be observed in the 3D NCI plots, but not in the 2D ones. Nevertheless, the NCI-ELMO method could be used as a tool to identify the presence of n-π^* interactions, which would be computationally less demanding than performing NBO analyses and, at the same time, more reliable than the previously used geometrical criteria.

9.2.3.4 Interactions with metal centers

As a final example, we investigated the interactions with a metal center in a protein. Metals are very common in proteins and are crucial for their functions and the stability of their structures.[381] For our study,[320] we chose the X-ray structure[444] (PDB code: 3UMI) of the amyloid precursor protein E2 domain, in which a Zn^{2+} cation is coordinated by three histidine residues (His382, His432 and His 436) and one water molecule. These interactions were studied with the promolecular-NCI and NCI-ELMO (cc-pVDZ basis set) strategies. Following the same procedure also applied in the benchmark computations on the HIV Zinc fingerlike domain, a spherical atomic electron distribution for the charged cation Zn^{2+} was used to compute the promolecular density. Additionally, the ELMOs for the Zn^{2+} ion were computed on a model molecule that mimicked the chemical environment of the metal in the chosen protein structure.

For each of the four interactions between the metal center and the environment, a disk-shaped RDG isosurface is observed in Figure 9.15A. Furthermore, in the 2D NCI plot (Figure 9.15B), each interaction corresponds to a specific spike in the negative region of the "signed" electron density. In contrast to the previously described interaction types, in this case the promolecular-NCI approach provides peaks occurring at less negative values of the "signed" electron density compared to the NCI-ELMO ones. This trend is fully consistent with the previous observations for the interactions of Zn^{2+} in the HIV Zinc fingerlike domain (compare Figure 9.8C).

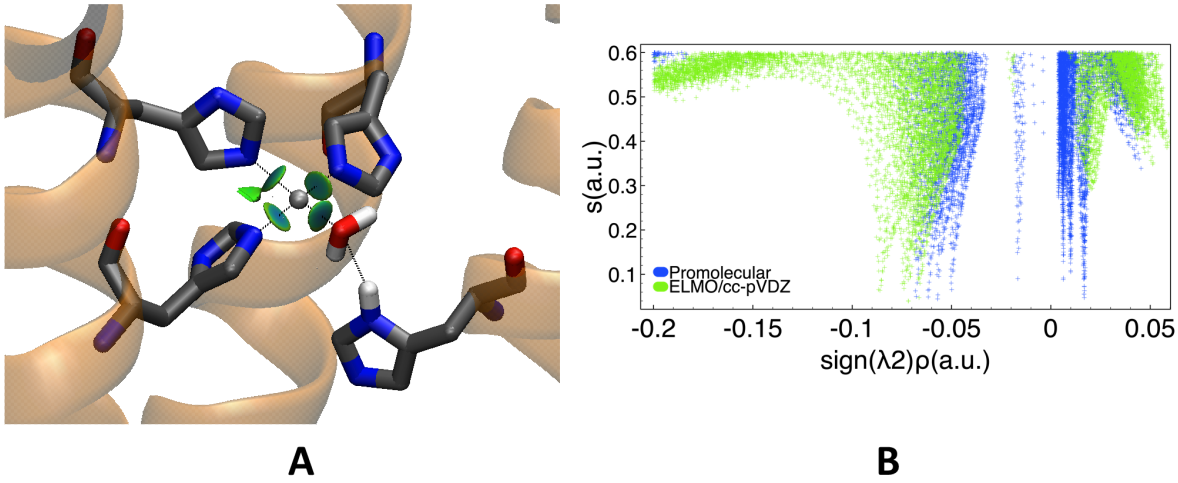


Figure 9.15: Interactions between the Zn^{2+} ion and coordinating residues (one water molecule and three histidine residues His382, His432 and His 436) in APP (PDB code: 3UMI): (A) RDG isosurfaces (s=0.6 a.u., color scale: -0.03 a.u. $< sign(\lambda_2)\,\rho <$ 0.03 a.u.) obtained at NCI-ELMO level; (B) 2D NCI plots obtained at promolecular-NCI and NCI-ELMO levels for basis set cc-pVDZ. Reprinted with permission from reference [320]. Copyright 2019 American Chemical Society.

9.2.4 Conclusions for the qualtiative analysis

The new NCI-ELMO method represents an important extension of the NCI technique because the analysis of non-covalent interactions in large molecules can now be based on the ELMO description, which fully takes into account the asphericity of the electron density. This represents an important step forward because the NCI analyses for such systems previously resorted to the crude promolecular approximation, which completely neglects the aspherical features of the electron distributions.

All the results presented in this section showed that the NCI-ELMO method is more reliable in classifying non-covalent interactions than the promolecular approximation. In fact, as indicated by the benchmark calculations, the new technique provides results that are in very good agreement with NCI analyses based on DFT calculations. However, since the computational cost of the NCI-ELMO technique is significantly lower compared to the NCI-DFT method, it allows the study of non-covalent interactions also in proteins, as shown in this section for a large variety of systems.

Therefore, as can be seen from the presented results, the new NCI-ELMO technique offers the possibility for rapid and reliable studies of non-covalent interactions in large biological systems.

9.3 Quantitative analysis

In this section, a possible extension of the NCI-ELMO technique to a more quantitative analysis of non-covalent interactions will be presented. This work is currently still ongoing. Therefore, the results presented in this part of the thesis are preliminary. In particular, we have systematically re-evaluated the parameters for the computation of promolecular-NCI integrals and we have newly defined them also for the calculation of NCI-ELMO integrals. The results of this parametrization will be presented in Section 9.3.1. Afterwards, we have compared the NCI results to those obtained from DFT calculations. This comparison will be shown in Section 9.3.2. Finally, we have applied the NCI-ELMO strategy to evaluate interactions between ligands and surrounding residues in proteins. The corresponding results will be discussed in Section 9.3.3.

9.3.1 Parametrization of the NCI integrals

Before the NCI integrals may be used to assess the strengths of non-covalent interactions, the variable parameters in Equations (9.3) and (9.4) need to be thoroughly determined for the different types of underlying electron densities. In the following, the adopted procedure to obtain these parameters will be described.

9.3.1.1 Benchmark interaction energies

The NCI integral values were parameterized against benchmark interaction energies. In particular, we used again the interaction energies in the S66 database.[396] As already mentioned in Section 9.1.2, this dataset consists of 66 small molecular dimers whose structures and interaction energies were computed at CCSD(T)/CBS level with the supermolecular approach and the counterpoise correction.

9.3.1.2 Underlying electron density calculations

As described in Section 9.1.2, fully QM densities are not suitable to compute the NCI integrals because they cannot be separated into individual monomer contributions according to Equation (9.2). Therefore, in this study, other than the usual promolecular densities, we also considered two types of densities resulting from the transfer of ELMOs.

To obtain the two types of ELMO densities, in a preliminary step the ELMOs for each of the 14 monomers in the S66 database were computed adopting the same procedure used to set up the ELMO libraries (compare Section 1.3.4) and exploiting all five basis sets of the ELMO databanks (namely, 6-31G, 6-311G, 6-31G(d,p), 6-311G(d,p) and cc-pVDZ). Afterwards, these ELMOs were included in the *ELMOdb* program and were transferred to the equilibrium geometries of the 66 dimers in the S66 dataset.

For the subsequent orthogonalization of the ELMOs (which is necessary for the electron density computation), two different procedures were applied. In the first case (henceforth called "dimer approximation"), all the ELMOs located on both monomers constituting a dimer were Löwdin orthogonalized among themselves, so that all the ELMOs in the dimer are mutually orthogonal to each other, but the electron distribution of each monomer is

slightly delocalized on the other monomer. Nevertheless, it should be mentioned that the Löwdin orthogonaliation procedure preserves the localization of the ELMOs to a large extent. Therefore, it was still possible to approximately separate the two monomer contributions.

However, to completely avoid any delocalization of the electron density or, in other words, to obtain two unambiguously separable electron densities, in the second case (hereafter called "monomer approximation"), only those ELMOs located on the same monomer were Löwdin orthogonalized among themselves. In this situation, the orbitals located on different monomers are not orthogonal to each other, but two strictly localized electron densities can be obtained for each monomer.

9.3.1.3 Parametrization

To compute the NCI integrals, the best combination of electron density exponent n, electron density threshold γ^{ref} and reduced density gradient threshold s^{ref} (compare Equations (9.3) and (9.4)) needs to be determined for each type of underlying density. In particular, for each type of electron distribution, we only redefined the first two values, while s^{ref} was kept fixed to 1.0. This choice was made[414] to include all the relevant interactions, as already explained in Section 9.1.2. For the definition of the electron density exponent n, a range of values between 1.0 and 3.0 with a stepsize of 0.1 was considered, while the electron density threshold γ^{ref} was increased from 0.51 to 0.99 using a stepsize of 0.01. For all the possible combinations of n and γ^{ref}, with a reduced density gradient thresholds $s^{ref} = 1.0$, the NCI integrals were computed for each of the dimers in the S66 database using Equation (9.4). Afterwards, the correlation between NCI integrals and interaction energies was evaluated, and the best combination was the one with the highest correlation coefficient. All the NCI integrals were calculated using a modified version of the *NCIPLOT4* software that allows the computation of NCI integrals also for fully QM densities and includes the option to evaluate the full range of exponents n and electron density thresholds γ^{ref}.

9.3.1.4 Results and discussion

In Figure 9.16, the resulting values of the correlation coefficients are depicted as a function of the electron density exponent n and electron density threshold γ^{ref} used to compute the NCI-integral values for the three different types of electron distributions. The results obtained with the promolecular densities are shown in Figure 9.16A, and indicate that the correlation coefficient is very much dependent on the choice of the parameters. In particular, high correlation coefficients (above 0.9), are only obtained for large values of n and γ^{ref}. In contrast, for lower values of these parameters, the interaction energies do not correlate with the promolecular-NCI-integral values. In contrast to these results, when the NCI integrals are computed using the ELMO electron densities with basis set 6-31G(d,p) (panels B and C in Figure 9.16), the correlation coefficients remain more stable and higher for all the different parameters. Analogous results were obtained for the other four sets of basis functions (compare heatmaps in Figures C.4-C.7 in the Appendix).

The observed differences can be explained by plotting the RDG isosurfaces using an isovalue of 1.0 (this corresponds to the adopted RDG threshold s^{ref}) and removing all the points where one of the monomer densities is larger than γ^{ref}. The remaining points included in

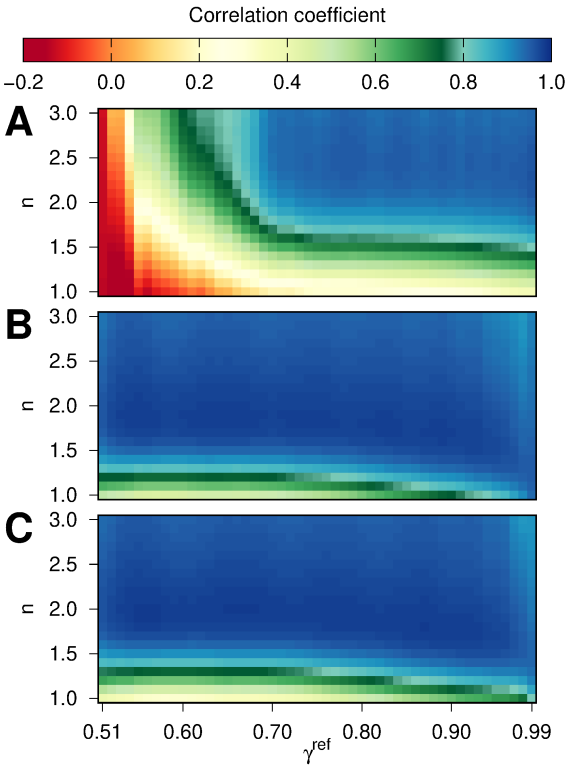


Figure 9.16: Heatmaps showing the variation of the correlation coefficients as a function of the electron density exponent n and electron density threshold γ^{ref}. Each correlation coefficient gives the correlation between the interaction energies and the NCI-integral values for the S66 dataset. The following densities were used to compute the NCI-integrals: (A) promolecular, (B) ELMO/6-31G(d,p) (monomer approximation) and (C) ELMO/6-31G(d,p) (dimer approximation).

the isosurfaces correspond to the integration domains Ω_{NCI} (also defined as intermolecular interaction regions in Section 9.1.2). For the example of the methylamine-methanol dimer, the integration domains are depicted in Figure 9.17 for three types of densities and different values of γ^{ref}. For the promolecular approximation, the size of the integration domain largely depends on the value of γ^{ref}. Interestingly, no integration domain remains for $\gamma^{ref} = 0.55$, which can explain the negative correlation coefficients observed for low values of γ^{ref}. In fact, for many of the studied dimers, the integration domains vanish completely if γ^{ref} is low. However, for the promolecular approximations, the integration domains become very large as γ^{ref} increases. Contrary to what is observed for the promolecular density, for both ELMO approximations, an integration domain can always be observed. Although also this region becomes larger when γ^{ref} increases, it overall remains quite stable.

These trends can be also noticed in the corresponding 2D NCI plots, which are shown in Figure 9.18. In fact, also in this case, the plots associated with the promolecular-NCI analysis change significantly when γ^{ref} increases, while the NCI-ELMO plots (basis set 6-31G(d,p), both approximations) are much less influenced by the choice of γ^{ref}.

Finally, in Table 9.2, the best combinations of electron density exponent n and electron density threshold γ^{ref} are reported for each of the exploited electron densities. The optimal parameters are those giving the highest correlation coefficients, which are also listed in

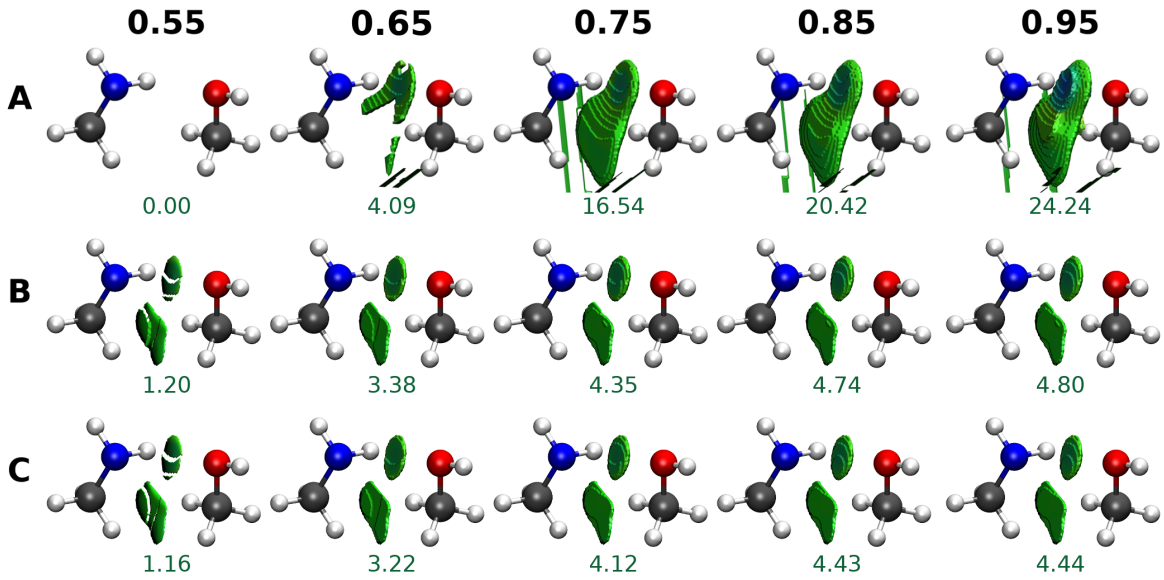


Figure 9.17: Size of the integration domains Ω_{NCI} associated with the intermolecular interactions in the methylamine-methanol dimer as a function of γ^{ref}, with $s^{ref} = 1.0$. The following densities were used in the NCI analyses: (A) promolecular, (B) ELMO/6-31G(d,p) (monomer approximation) and (C) ELMO/6-31G(d,p) (dimer approximation). Below each isosurface, the volume (in a.u.) of the integration domain Ω_{NCI} is also given.

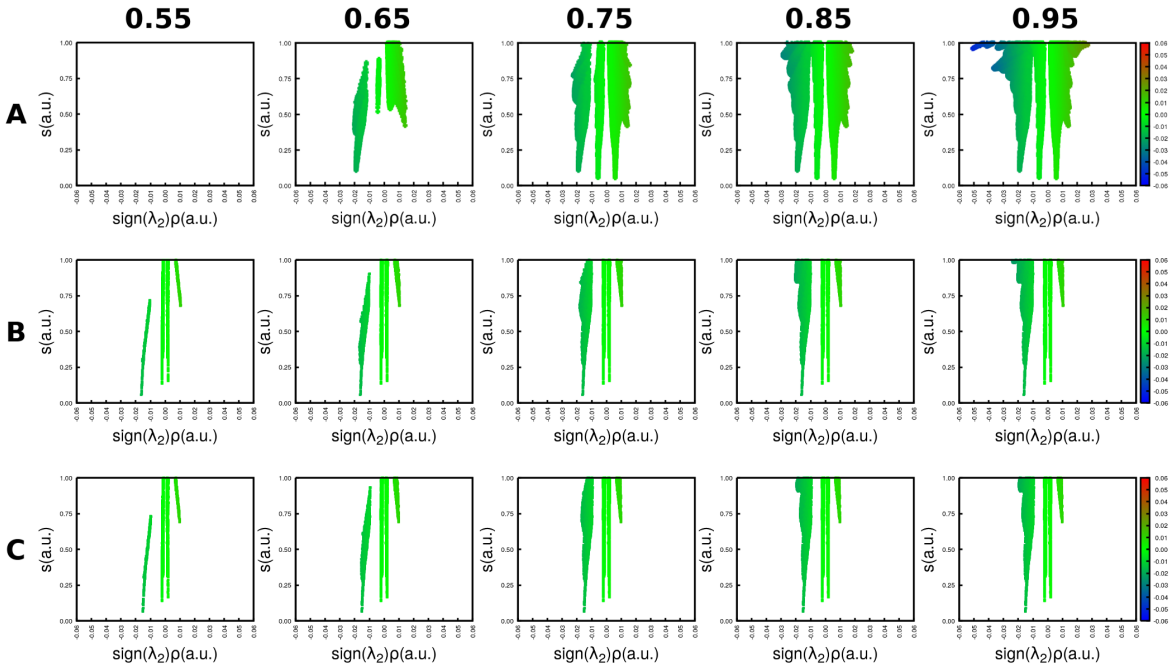


Figure 9.18: 2D NCI plots for the intermolecular interactions in the methylamine-methanol dimer as a function of γ^{ref}. The following densities were used in the NCI analyses: (A) promolecular, (B) ELMO/6-31G(d,p) (monomer approximation) and (C) ELMO/6-31G(d,p) (dimer approximation).

Table 9.2. Comparing these values among the different electron density methods, it can be concluded that the best correlation with the interaction energies is always obtained for the NCI-ELMO integrals, for which the correlation coefficients lie between 0.973 and 0.983. For the promolecular-NCI integrals, the previously recommended parameters were $n = 2.5$ and $\gamma^{ref} = 0.95$ (see Section 9.1.2).[414] These values were also determined for the S66 database. However, in contrast to this previous study, the new parameters (namely, $n = 2.3$ and $\gamma^{ref} = 0.99$) were obtained more systematically by using a smaller stepsize for n and γ^{ref}. Nevertheless, the previously recommended parameters are actually very close to the ones that gave the best correlation in this study.

Concerning the two different approximations to obtain the ELMO densities, the dimer approximation always yields slightly higher correlation coefficients than the monomer approximation. Therefore, especially for small systems, the orbitals should always be completely orthogonalized. However, the monomer approximation could still be helpful for the study of larger systems since the associated computational cost is lower when fewer orbitals need to be orthogonalized. Moreover, the monomer approximation could be also particularly advantageous from the computational point of view if one has to evaluate the interactions between a protein and several ligands. In this case, the orthogonalization of the ELMOs associated with a protein can be performed only once, while only the orthogonalizations of the fewer ELMOs corresponding to the ligands need to be performed for each evaluation of protein-ligand interaction.

In the remaining part of this chapter, the combination of parameters reported in Table 9.2 will be always used for the computation of NCI-integral values.

Table 9.2: Combinations of electron density exponent n and electron density threshold γ^{ref} leading to the highest correlation coefficients for the correlation between NCI-ELMO integral values and interaction energies (S66 dataset). The values for the promolecular density are also reported in the last row of the table.

Electron density	Monomer approximation			Dimer approximation		
	Corr. coeff.	n	γ^{ref}	Corr. coeff.	n	γ^{ref}
ELMO/6-31G	0.974	1.8	0.60	0.975	1.9	0.56
ELMO/6-311G	0.973	1.9	0.63	0.975	2.0	0.65
ELMO/6-31G(d,p)	0.981	1.9	0.55	0.983	2.0	0.56
ELMO/6-311G(d,p)	0.978	2.0	0.58	0.981	2.1	0.59
ELMO/cc-pVDZ	0.974	2.0	0.57	0.976	2.1	0.58
Promolecular	0.958	2.3	0.99			

9.3.2 Comparison to DFT calculations

To better contextualize the results obtained in the previous subsection for the correlation between the NCI integrals and interaction energies, the correlation coefficients can be compared to those resulting from traditional quantum chemical calculations. In particular, interaction energies were computed with the supermolecular approach (compare Section 8.2.2) at DFT

level using the PBE functional[33] with and without Grimme's D3 dispersion correction.[395] All calculations were performed with counterpoise correction[397] using *Gaussian09*.[150] In combination with PBE and PBE-D3, the following sets of basis functions were considered: the five basis sets used in the ELMO database plus 6-31+G(d,p),[136,137,144,445] 6-311+G(d,p)[141,445] and aug-cc-pVDZ.[145,446] As done for the NCI integrals, the resulting energies were compared to the S66 benchmark values, which were computed at CCSD(T)/CBS level. The obtained correlation coefficients, mean absolute differences and mean absolute percentage errors are listed in Table 9.3.

Table 9.3: Correlation coefficients, mean absolute differences (MADs) and mean absolute percentage errors (MAPEs) computed for the promolecular-NCI, NCI-ELMO (monomer and dimer approximation), PBE and PBE-D3 calculations with respect to the CCSD(T)/CBS benchmark values in the S66 database.

Method/ basis set	Corr. coeff.	MAD (kcal/mol)	MAPE (%)
promolecular	0.958	0.84	18.4%
ELMO/6-31G (monomer)	0.974	0.69	15.8%
ELMO/6-311G (monomer)	0.973	0.68	15.4%
ELMO/6-31G(d,p) (monomer)	0.981	0.63	14.7%
ELMO/6-311G(d,p) (monomer)	0.978	0.65	15.0%
ELMO/cc-pVDZ (monomer)	0.974	0.69	16.1%
ELMO/6-31G (dimer)	0.975	0.67	15.0%
ELMO/6-311G (dimer)	0.975	0.64	14.5%
ELMO/6-31G(d,p) (dimer)	0.983	0.57	13.3%
ELMO/6-311G(d,p) (dimer)	0.981	0.59	13.8%
ELMO/cc-pVDZ (dimer)	0.976	0.66	15.3%
PBE/6-31G	0.889	2.55	65.2%
PBE/6-311G	0.897	2.40	61.9%
PBE/6-31G(d,p)	0.910	2.22	59.8%
PBE/6-311G(d,p)	0.920	2.14	56.7%
PBE/6-31+G(d,p)	0.918	2.11	56.3%
PBE/6-311+G(d,p)	0.921	2.10	55.5%
PBE/cc-pVDZ	0.920	2.35	61.1%
PBE/aug-cc-pVDZ	0.929	2.16	57.1%
PBE-D3/6-31G	0.958	1.35	24.1%
PBE-D3/6-311G	0.965	1.18	20.1%
PBE-D3/6-31G(d,p)	0.976	0.79	15.8%
PBE-D3/6-311G(d,p)	0.983	0.58	11.7%
PBE-D3/6-31+G(d,p)	0.983	0.62	11.8%
PBE-D3/6-311+G(d,p)	0.985	0.57	10.7%
PBE-D3/cc-pVDZ	0.983	0.45	9.0%
PBE-D3/aug-cc-pVDZ	0.987	0.44	8.0%

Let us first focus on the promolecular-NCI and NCI-ELMO results. As already described above for the correlation coefficients in the previous subsection, also the mean absolute differences and mean absolute percentage errors are always better for NCI-ELMO than for promolecular-NCI. When ELMO densities are used, the values are slightly higher for the monomer approximation than for the dimer one. Furthermore, all the NCI analyses result in mean absolute differences that are lower than the chemical accuracy limit (1.0 kcal/mol).

However, when compared to the PBE results, all the NCI analyses provide results that are in much better agreement with the benchmark values, regardless of the chosen basis set. In fact, the mean absolute difference between the PBE results and the benchmark ones is always more than 2 kcal/mol. Better agreement can be obtained when the dispersion correction is introduced, namely when an extra term describing the dispersion contribution is added to the total DFT energy. By introducing this correction, the correlation coefficients increase, while the differences and errors are significantly reduced. Nevertheless, for the two Pople basis sets without polarization and diffuse functions (6-31G and 6-311G), the results are still worse than those obtained through the NCI analyses. In contrast, when polarization and diffuse basis functions are used, the results with PBE-D3 are similar to the NCI-ELMO ones. Even better agreement with the benchmark values can be obtained for PBE-D3 calculations with the correlation consistent basis sets cc-pVDZ and aug-cc-pVDZ.

It is possible that the NCI results could be further improved if more relaxed densities would be used. One way of achieving this goal could consist in polarizing the ELMOs, which is a development currently envisaged.

In conclusion, we have used the best set of parameters to compute the NCI integrals and we have compared the results to those resulting from DFT calculations. These comparisons were based on correlation coefficients, mean absolute differences and mean absolute percentage errors, with the S66 interaction energies computed at CCSD(T)/CBS level as references. The performed calculations showed that the novel NCI-ELMO integral values can provide interaction energies that agree with the benchmark values as well as PBE-D3 calculations. For this reason, we decided to try to exploit the newly parameterized NCI-ELMO integrals to evaluate interactions in larger systems of biological interests (see next subsection of the chapter).

9.3.3 Quantitative NCI analyses applied to protein-ligand complexes

In the last part of this chapter, the novel NCI-ELMO integrals will be applied to two different protein-ligand complexes. The non-covalent interactions between proteins and ligands are of particular interest because they are crucial for molecular recognition processes. In this context, it can be particularly helpful to know how strongly each residue in the protein is interacting with the ligand.

However, the ligands in such complexes often require a more thorough quantum mechanical description than the one offered by the ELMO technique. For example, if the ligand is a chromophore, the ELMOs are not sufficient to describe this highly delocalized system. In this regard, the QM/ELMO technique (compare Section 1.5.1) is particularly useful as it provides a more accurate density for the ligand. Therefore, the different protein-ligand complexes have been studied with the combined NCI-QM/ELMO approach. It is worth noting that the use of the QM/ELMO approach still guarantees the separability of the electron densities associated with the interacting units (in this particular case, the protein and the ligand), which is the fundamental working hypothesis at the basis of the quantitative NCI strategy.

9.3.3.1 Model systems for the protein-ligand complexes

The work presented in this part of the chapter is currently still ongoing. As first two examples, we have chosen to study the following protein-ligand complexes: (i) the green fluorescent protein with its chromophore (p-hydroxybenzylidene-imidazolinone, pHBDI) and (ii) the CFTR associated ligand (CAL) PDZ domain with the polypeptide iCAL36. The preliminary results for these two protein-ligand complexes will be discussed in the remaining part of this chapter.

The structure of each ligand and its interactions with the surrounding protein residues are schematically represented in 2D using *LIGPLOT*[447,448] diagrams. These were obtained with the program *LigPlot+*,[448] which generates a 2D diagram of the protein-ligand complex and which uses the software *HBPLUS*[377] to detect the interactions between ligand and protein. This is done by using geometrical criteria. In the resulting diagrams, hydrogen bonds are depicted as green dashed lines, while non-bonded interactions are shown as red arcs with spokes pointing towards the interacting ligand atoms. The *LIGPLOT* diagrams of the protein-ligand complexes are shown in Figures 9.19a and 9.21a for systems (i) and (ii), respectively.

For all the protein-ligand complexes, reduced model systems were used for the NCI analyses. The model system for the green florescent protein was extracted from the crystal structure (PDB code: 1EMB)[449] by Kaila and co-workers.[450] It consists of the chromophore pHBDI, nine surrounding residues (Thr62, Gln69, Gln94, Arg96, His148, Val150, Thr203, Ser205 and Glu222) and four water molecules. Since the chromophore is bonded to the protein, the backbone atoms of Phe64 (C, O and carbon α) and Val68 (N and carbon α) are also part of the model system. The nine surrounding amino acids were cut at the carbon β atom. Kaila and co-workers also optimized the structure of this model system at B3LYP/def2-SVP level keeping fixed the carbon β atoms of the nine surrounding amino acids and the two carbon α atoms of Phe64 and Val68.[450]

For the second protein-ligand complex, the model system was initially prepared for the first QM/ELMO study.[205] In a first step, chain B and D were selected from the crystal structure (PDB code: 4E34)[451] for the PDZ domain and the polypeptide, respectively. Afterwards, the model system was extracted from the two chains, considering the polypeptide iCAL36 and the surrounding residues (those within a 5 Å radius around the ligand). The cut bonds were saturated using N-methyl amino (CH_3-NH-) and acetyl (CH_3-CO-) groups. The hydrogen atoms were added to the model system according to the pH value using *LEaP* in *AMBER*[148] and were afterwards optimized at HF/cc-pVDZ level.[205] The final model system consists of 47 residues, including two water molecules, and 550 atoms.

9.3.3.2 QM/ELMO calculations and subsequent NCI analysis

All the QM/ELMO calculations were performed with the ligand in the QM region and the surrounding protein in the ELMO region using the M06-2X functional[452] for the QM part and the basis set cc-pVDZ for the whole system. All QM/ELMO calculations were carried out with our in-house modified version of *Gaussian09*[150] (for more details about the QM/ELMO technique, see Section 1.5.1).

In the subsequent NCI analyses, the individual interactions between the ligand and each of the surrounding protein residues were computed separately. This is possible since the QM or-

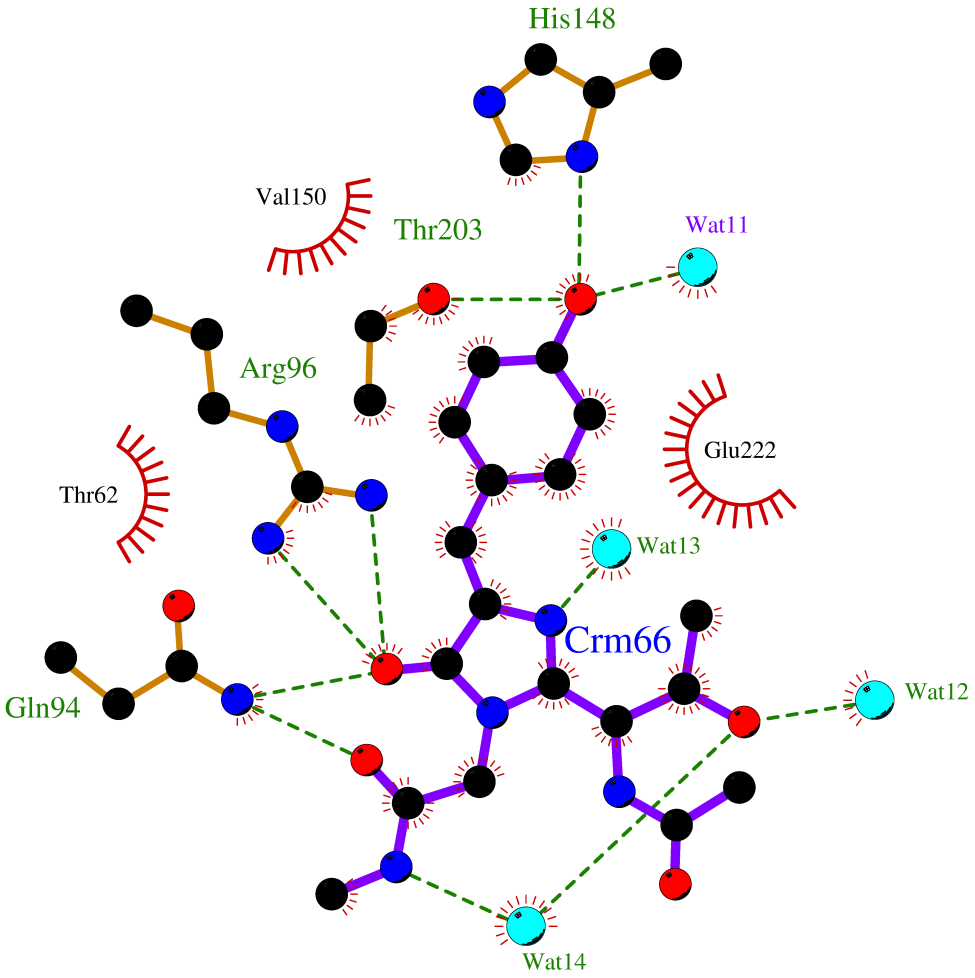
bitals remain mostly localized in the QM region, while the surrounding residues are described by means of ELMOs. Therefore, to separate the different contributions, the previously described dimer approximation was applied (see Section 9.3.1). Hence, the NCI integrals were computed using the optimal set of parameters obtained for ELMO/cc-pVDZ with the dimer approximation as reported in Table 9.2. Furthermore, to also assess the contributions of different interaction types, the NCI integrals were evaluated considering two subdomains of integration according to the value of the "signed" electron density ρ. In particular, values of $sign(\lambda_2)\rho$ between -0.1 and -0.02 a.u. have been assigned to attractive interactions and those between -0.02 and 0.02 to van der Waals interactions.[414] All NCI analyses were performed using the previously mentioned modified version of *NCIPLOT4*.[414]

9.3.3.3 Results for the green fluorescent protein

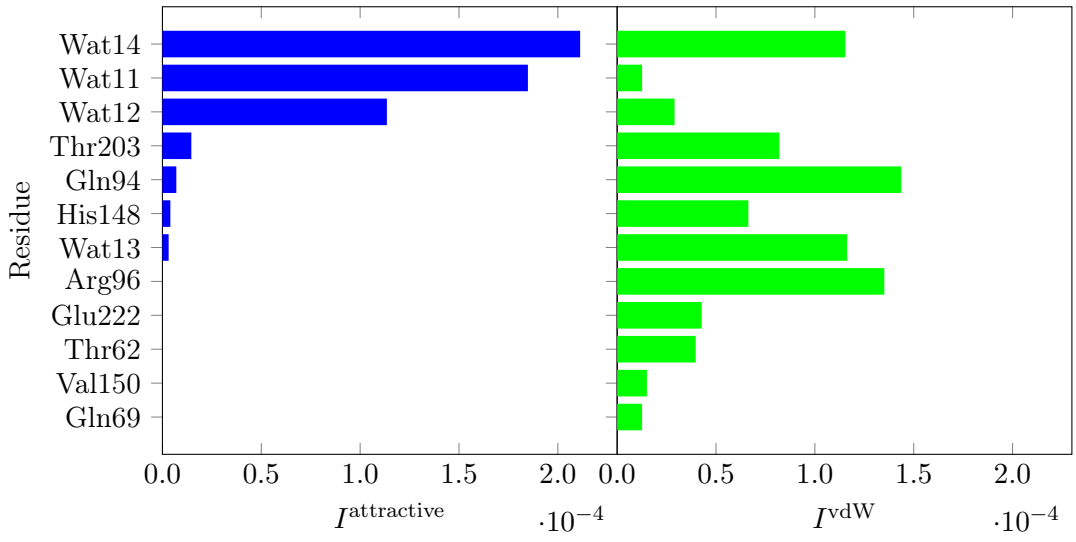
The green fluorescent protein has become an indispensable tool in biology and medicine.[453] It is responsible for green fluorescence in a certain type of jellyfish, *Aequorea aequorea*,[453] and is nowadays used as a marker for gene expression and for targeting proteins in cells and organisms.[454]

Exploiting the NCI method based on QM/ELMO densities, we have studied the interactions between the chromophore and the surrounding residues in the model for the green fluorescent protein as described above. As can be seen in Figure 9.19a, the chromophore could be involved in several hydrogen bonds and non-bonded interactions. The aim of the study is to better clarify the nature of the different interactions. Therefore, in Figure 9.19b, the corresponding NCI integral values are shown separately for the attractive ($I^{\text{attractive}}$) and the van der Waals (I^{vdW}) contributions. Furthermore, the integration domains associated with the intermolecular interactions between the chromophore and each individual residue are depicted in Figure 9.20. Finally, the NCI integral values are also reported in Table 9.4, together with the hydrogen-acceptor distances for possible hydrogen bonds.

The largest integral values in the "attractive" integration domain in Figure 9.19b and Table 9.4 are those corresponding to the interactions of the water molecules Wat14, Wat11 and Wat12 with the chromophore. In Figures 9.20a and 9.20b, disk-shaped integration domains with blue centers can be observed between all those three water molecules and the chromophore, indicating that these interactions are indeed hydrogen bonds. It is interesting to note that, among the three water molecules, the largest integral values in both integration domains are obtained for Wat14, which is involved in two hydrogen bonds with the chromophore, as indicated by the two disk-shaped integration domains in Figure 9.20b. Furthermore, in the same plot, a large green integration domain is located below Wat14. Therefore, it can be expected that both hydrogen bonds are contributing to the large attractive integral value, while the large van der Waals integral is caused by the other detected interaction. For Wat11 and Wat12, the attractive integral values are slightly lower than for Wat14, which can be explained with the fact that each of the water molecules forms only one hydrogen bond with the chromophore. However, these hydrogen bonds are rather short (compare Table 9.4), nevertheless leading to large attractive contributions. Concerning the difference between Wat11 and Wat12, a higher attractive integral value is observed for Wat11, which is indeed the one involved in a shorter hydrogen bond. Furthermore, for Wat11 and Wat12, small integral val-

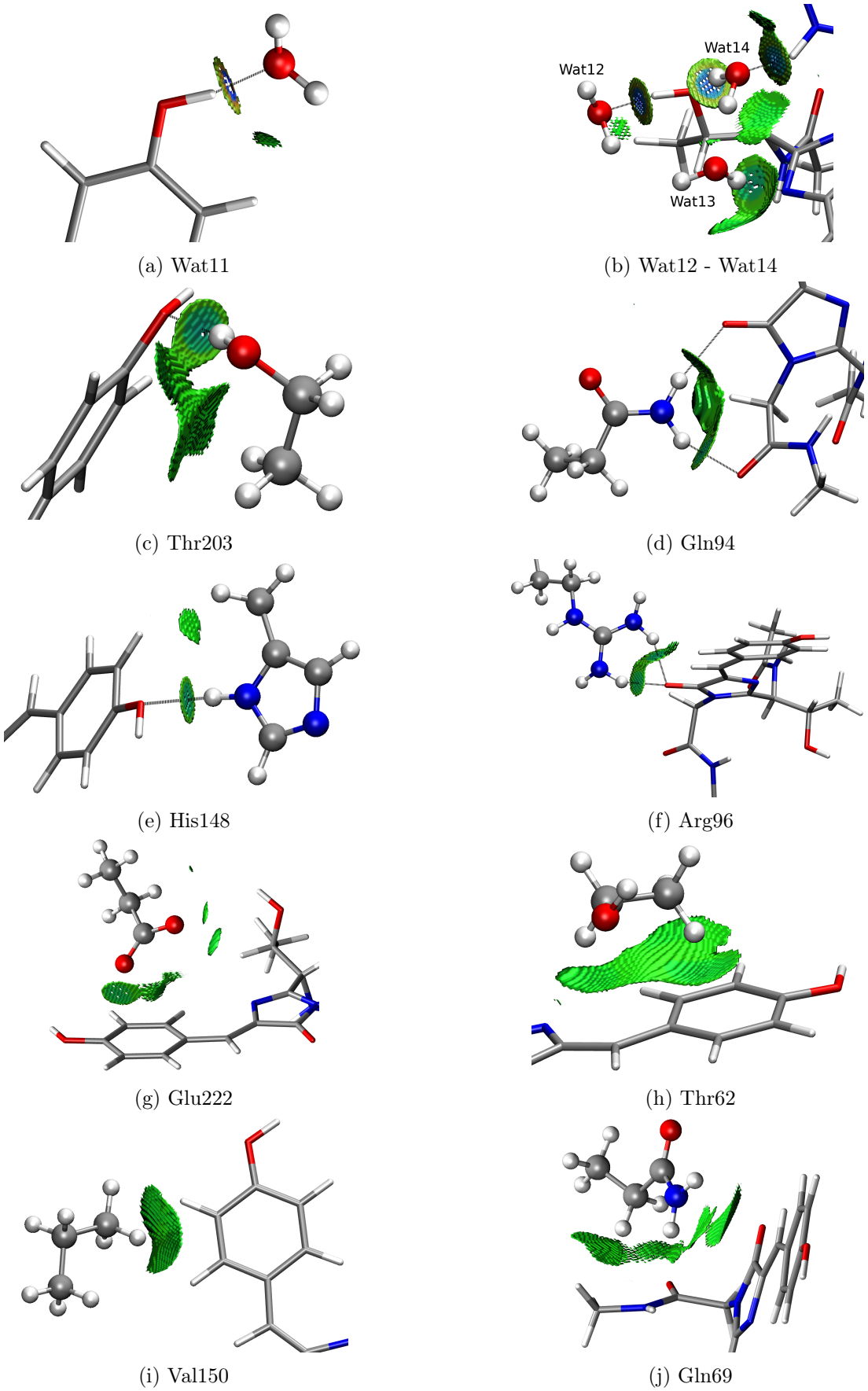


(a) *LIGPLOT*



(b) NCI integral values

Figure 9.19: (a) *LIGPLOT* diagram showing the interactions between the chromophore and the surrounding residues in the green fluorescent protein. The green dotted lines indicate hydrogen bonds and the red arcs represent those residues that are expected to form non-bonded contacts with the ligand. (b) NCI integral values for the interaction of each residue with the chromophore in the green fluorescent protein.



(a) Wat11

(b) Wat12 - Wat14

(c) Thr203

(d) Gln94

(e) His148

(f) Arg96

(g) Glu222

(h) Thr62

(i) Val150

(j) Gln69

Figure 9.20: Integration domains Ω_{NCI} associated with the intermolecular interactions of the indicated residue (in ball and stick representation) with the chromophore (in licorice representation) in the green fluorescent protein.

Table 9.4: NCI integral values for the interactions between the chromophore and the surrounding residues belonging to the green fluorescent protein. For residues that could be involved in hydrogen bonds with the chromophore, the distances between the hydrogen and the acceptor atoms are also listed. More than one value indicates that the residue is possibly involved in more than one hydrogen bond.

Residue	Integral values ($\cdot 10^{-4}$)		H$\cdots$A distance(s) (in Å)
	$I^{\text{attractive}}$	I^{vdW}	
Wat14	2.11	1.15	1.725, 1.797
Wat11	1.85	0.13	1.517
Wat12	1.14	0.29	1.612
Thr203	0.15	0.82	1.937
Gln94	0.07	1.44	1.985, 2.140
His148	0.04	0.66	1.963
Wat13	0.03	1.16	1.995
Arg96	0.00	1.35	2.040, 2.145
Glu222	0.00	0.43	
Thr62	0.00	0.40	
Val150	0.00	0.15	
Gln69	0.00	0.13	

ues are also obtained in the "van der Waals" integration domain, which probably correspond to the small green isosurfaces below Wat11 in Figure 9.20a and to the one behind Wat12 in Figure 9.20b.

In addition to the hydrogen bonds between the water molecules and the chromophore, at least in principle further hydrogen bonds could exist between the chromophore and residues Thr203, Gln94, His148, Wat13 and Arg96, as indicated in the *LIGPLOT* diagram in Figure 9.19a. For the first four residues, significantly smaller attractive NCI integral values are obtained compared to the water molecules Wat14, Wat11 and Wat12. Furthermore, in all four cases, the attractive NCI integral values are significantly smaller than the corresponding values in the "van der Waals" integration domain. This can be explained with the fact that the hydrogen-acceptor distances are significantly longer than the corresponding distances observed for the hydrogen bonds established by Wat14, Wat11 and Wat12 (see Table 9.4). For the fifth residue, namely Arg96, the quantitative NCI analysis indicates that the interactions with the chromophore is purely of van der Waals type. In fact, in this case, no attractive contribution is observed, which can be explained again with the fact that the corresponding hydrogen-acceptor distances are quite large (i.e., larger than 2.0 Å; see Table 9.4). Moreover, except for His148, all the integration domains correspond to large delocalized isosurfaces (compare Figures 9.20c-f), also indicating that these are van der Waals interactions.

As can be seen in the *LIGPLOT* diagram in Figure 9.19a and in Figure 9.20, residues Glu222, Thr62, Val150 and Gln69 are oriented in such a way that they form only non-bonded interactions with the chromophore. The corresponding interaction domains in Figures 9.20g-j are again quite delocalized, indicating that the interactions are of van der Waals type. This is indeed confirmed by the quantitative NCI analysis. In fact, the integral values obtained for Glu222, Thr62, Val150 and Gln69 are smaller than for all the other residues in the green fluorescent protein.

Finally, it is interesting to observe that, out of all considered residues, the largest NCI integral values in the "van der Waals" integration domain are obtained for Gln94 and Arg96. This can be rationalized by considering that both residues establish double interactions with the ligand. Furthermore, it is important to note that the sizes of the integration domains in Figure 9.20 do not correlate with the NCI integral values. For example in the case of Glu222, the sum of the integration domains (Figure 9.20g) is significantly smaller than the domain obtained for Thr62 (Figure 9.20h), despite the integral values are almost equal for these two interactions.

9.3.3.4 Results for the CAL PDZ domain in complex with iCAL36

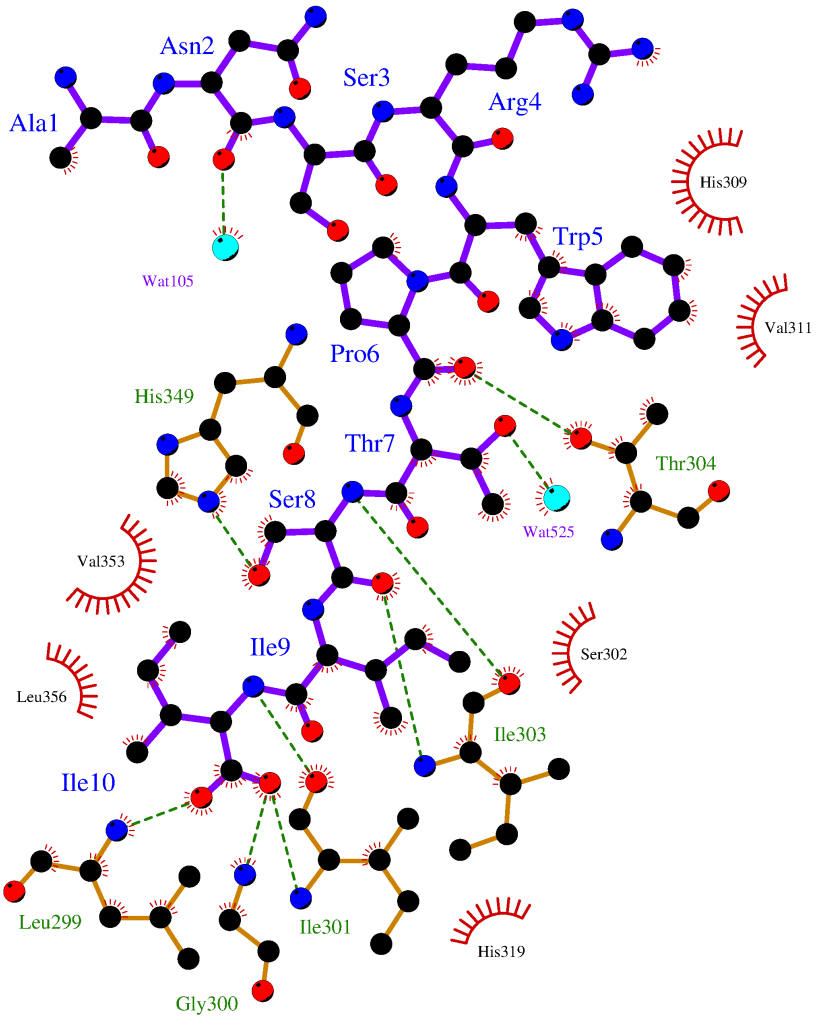
PDZ domains are one of the largest protein families of peptide recognition domains, which play an important role in many biological processes and in a number of diseases.[451] One potential therapeutic target for the PDZ domain in the cystic fibrosis transmembrane conductance regulator (CFTR) is the inhibitor peptide iCAL36. An analysis of the crystal structure of the protein and the inhibitor peptide (PDB code: 4E34) showed that mainly the six residues Trp5 to Ile10 of the peptide interact with the groove at the surface of the domain and that especially residue Ile10 of the peptide iCAL36 is important for the selectivity of the ligand-protein interaction.[451]

Here, we have analyzed the interactions between iCAL36 and the PDZ domain using the NCI method based on QM/ELMO densities. In Figure 9.21, the corresponding *LIGPLOT* diagram and the resulting NCI integral values are reported. Additionally, for each intermolecular interaction between the ligand and a particular surrounding residue, the integration domains are shown in Figure 9.22. Moreover, the actual integral values are listed in Table 9.5 together with the distances between hydrogen and acceptor atoms for possible hydrogen bonds.

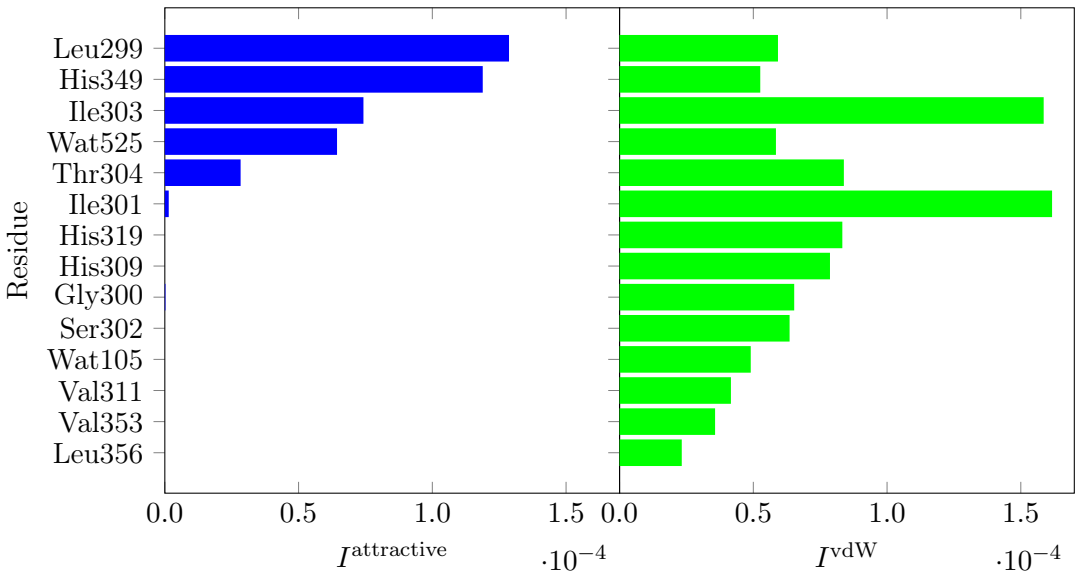
In comparison to the previously analyzed green fluorescent protein, where in most of the cases only one integration domain was observed for each considered residue (see Figure 9.20), in the case of the complex between the PDZ domain and the ligand iCAL, several integration domains are obtained for each considered residue (see Figure 9.22). This means that each of the considered residues in the PDZ domain is involved in several interactions with the ligand. In the following, the residues involved in attractive interactions will be analyzed first, followed by a discussion about the remaining residues.

In the quantitative NCI analysis, attractive integral values are obtained for the six residues Leu299, His349, Ile303, Wat525, Thr304 and Ile301 (see Figure 9.21b and Table 9.5). The integration domains for these six residues are reported in Figures 9.22a to 9.22f. In all six cases, at least one blue disk-shaped integration domain can be observed between groups that are typically involved in hydrogen bonds (as indicated by the black dotted lines), thus confirming that these residues are indeed involved in at least one hydrogen bond with the ligand.

Let us first focus on the residues that form one hydrogen bond with the ligand iCAL36 (Leu299, His349, Wat525 and Thr304) in the *LIGPLOT* diagram in Figure 9.21a. For each of these residues one blue disk-shaped integration domain is obtained. This is consistent with the significant NCI integral values reported in Figure 9.21b and Table 9.5 obtained for all four residues. Interestingly, the integral values decrease as the associated hydrogen-acceptor distances (Table 9.5) become larger.

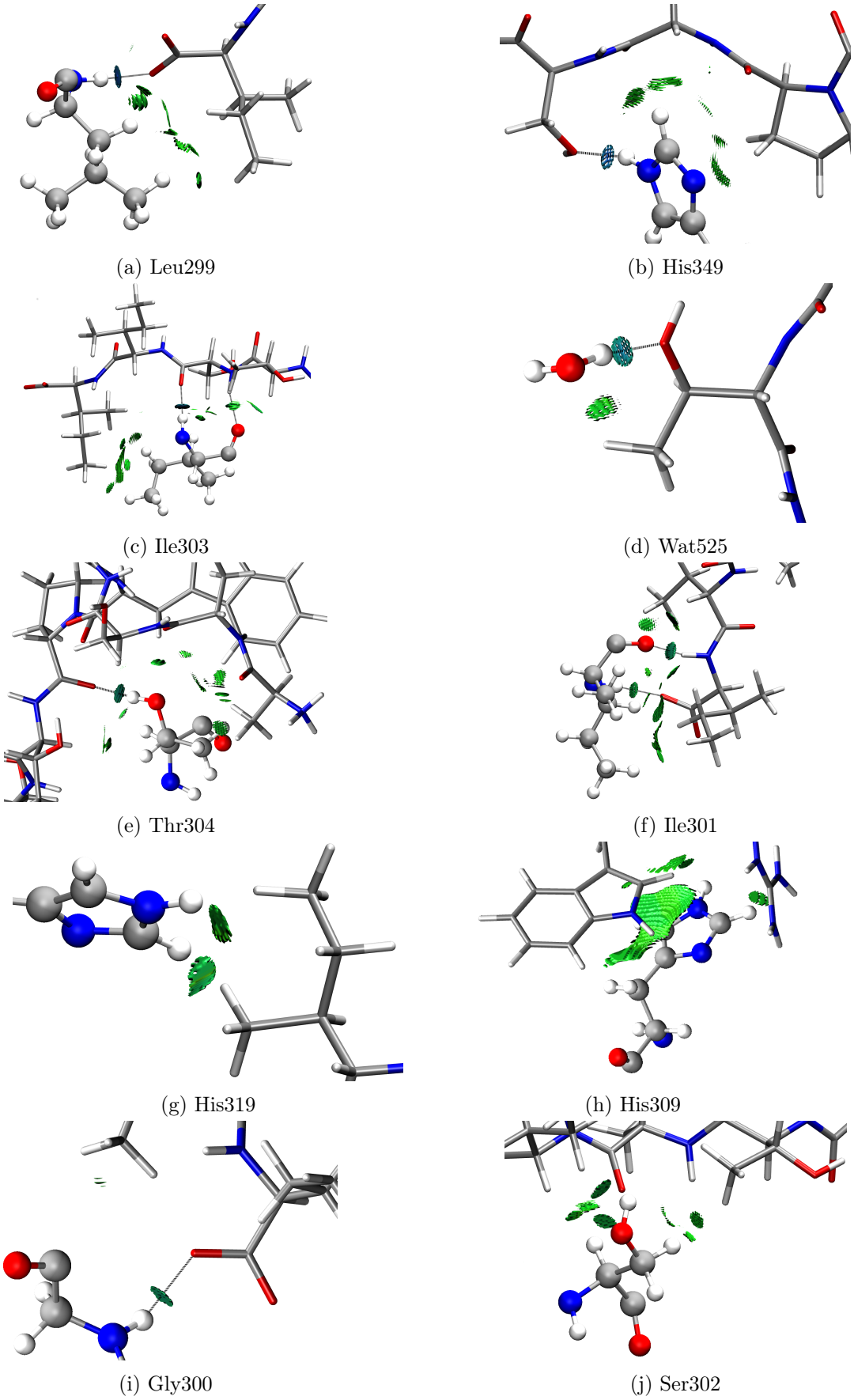


(a) *LIGPLOT*



(b) NCI integral values

Figure 9.21: (a) *LIGPLOT* diagram showing the interactions between the ligand iCAL36 and the surrounding residues in the PDZ domain. The green dotted lines indicate hydrogen bonds and the red arcs represent those residues that are expected to form non-bonded contacts with the ligand. (b) NCI integral values for the interaction of each residue with the ligand iCAL36.



(a) Leu299 (b) His349

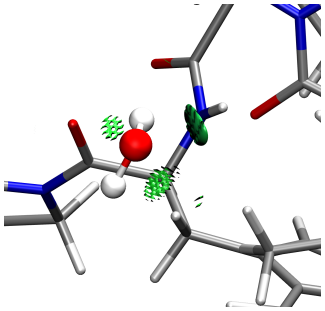
(c) Ile303 (d) Wat525

(e) Thr304 (f) Ile301

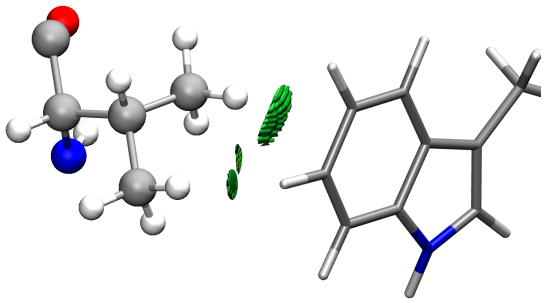
(g) His319 (h) His309

(i) Gly300 (j) Ser302

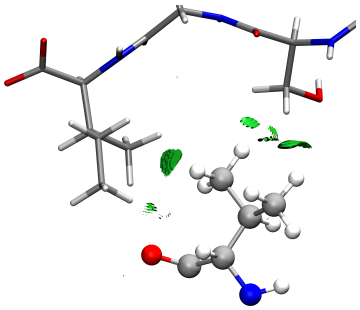
Figure 9.22: Integration domains Ω_{NCI} associated with the intermolecular interactions of the indicated residue in the PDZ domain (in ball and stick representation) with the ligand iCAL36 (in licorice representation). *Continued on next page.*



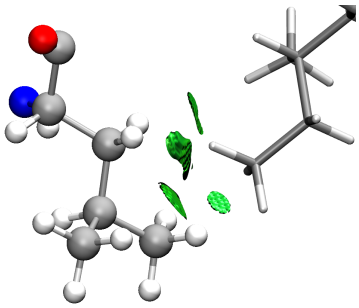
(k) Wat105



(l) Val311



(m) Val353



(n) Leu356

Figure 9.22: *Continued from previous page.*

In contrast to the previously discussed residues, both Ile301 and Ile303 could form two hydrogen bonds with the ligand iCAL36, as indicated by the dotted green lines in the *LIGPLOT* diagram. For both residues, corresponding integration domains are also observed in Figures 9.22c and 9.22f. However, the attractive integral value obtained for Ile303 is significantly larger than the one associated with Ile301. This can be explained with the distances between the hydrogen and acceptor atoms in Table 9.5. In particular, one short distance (1.758 Å) and a long one (1.984 Å) are obtained for Ile303, whereas two long distances (1.954 Å and 2.197 Å) are observed for Ile301. Therefore, the contacts associated with a distance greater than 1.9 Å are probably not part of the "attractive" integration domain, but of the "van der Waals" domain.

To obtain the preliminary NCI integral values presented in this section, we used the same integration domains as suggested in reference [414]. In particular, $sign(\lambda_2)\rho$ values lower than -0.02 a.u. were assigned to the attractive contributions, while values between -0.02 and 0.02 a.u. were associated with van der Waals interactions. However, these integration domains were proposed for promolecular densities and are not optimized for ELMO densities. The results of the qualitative analysis of non-covalent interactions presented in Section 9.2 suggest that such an optimization is probably necessary. In particular, in all the 2D NCI plots associated with hydrogen bonds, the NCI-ELMO peaks are shifted towards lower values of the signed electron density compared to the corresponding promolecular cases. Therefore, in the future, a new definition of the integration domains is certainly envisaged. A further possibility would be to introduce an intermediate category of weak attractive interactions, which would be associated with $sign(\lambda_2)\rho$ values around -0.02 a.u.

As already mentioned, all the considered residues in the PDZ domain are involved in several interactions with the ligand iCAL36. In fact, in Figure 9.22, quite delocalized green-colored interaction domains are observed for practically all the residues. This is consistent

Table 9.5: NCI integral values for the interactions between the iCAL36 ligand and the surrounding residues of the PDZ domain. For residues that could be involved in hydrogen bonds with the ligand, the distances between the hydrogen and the acceptor atoms are also listed. More than one value indicates that the residue is possibly involved in more than one hydrogen bond.

Residue	Integral values ($\cdot 10^{-4}$)		H$\cdots$A distance(s) (in Å)
	$I^{\text{attractive}}$	I^{vdW}	
Leu299	1.29	0.59	1.734
His349	1.19	0.53	1.749
Ile303	0.74	1.59	1.755, 1.984
Wat525	0.64	0.58	1.835
Thr304	0.28	0.84	1.877
Ile301	0.01	1.62	1.954, 2.197
His319	0.00	0.83	
His309	0.00	0.79	
Gly300	0.00	0.65	2.020
Ser302	0.00	0.64	
Wat105	0.00	0.49	
Val311	0.00	0.42	
Val353	0.00	0.36	
Leu356	0.00	0.23	

with the significant van der Waals NCI integral values observed for all residues (Figure 9.21b and Table 9.5). It is interesting to note that the largest NCI integrals in the "van der Waals" integration domain are obtained for the two isoleucine residues (Ile303 and Ile301). For these two residues, several integration domains can be observed in Figures 9.22c and 9.22f.

As previously noticed for the green fluorescent protein, the sizes of the integration domains do not correlate with the NCI integral values. For example, the interactions between the ligand and residues His319 or His309 are both characterized by almost equal van der Waals integral values, but the integration domain for His309 is significantly larger than the ones associated with His319.

Finally, only the first five residues of the ligand iCAL36 (Pro6 to Ile10) are involved in attractive interactions between protein and ligand, as can be evinced from the *LIGPLOT* diagram in Figure 9.21a and Table 9.5. This observation is in line with the observations by Amacher *et al.*, who identified the interactions of Ile10 as crucial for the selective binding of the ligand to the PDZ domain.[451] In fact, according to the *LIGPLOT* diagram and the integration domains, the PDZ domain is involved in several interactions with residues Leu299, Gly300, Ile301, Val353 and Leu356. Among these is also the strongest attractive interaction between the ligand and the protein, namely the hydrogen bond formed between Leu299 and Ile10.

In the future, it could be interesting to compute the ELMO density also of the ligand, and to perform a quantitative NCI analysis to estimate the strength of the interactions between the individual residues in the ligand and the surrounding PDZ domain. In this way, the interaction formed by the ten residues in the ligand could be ranked according to their strength, which would also allow us to further investigate the role of the Ile10 residue.

9.3.4 Conclusions for the quantitative analysis

With the goal of extracting quantitative information about non-covalent interactions through the NCI method, we have extended the novel NCI-ELMO strategy exploiting the calculation of the so-called NCI-ELMO integrals. These integrals were parameterized against benchmark interaction energies and applied to protein-ligand interactions.

From the parametrization of the NCI integral values against the benchmark interaction energies in the S66 database, it clearly emerged that the new NCI-ELMO method gives better correlations with the reference values than the previously used promolecular-NCI strategy. Furthermore, the integrals provided by the NCI-ELMO technique remain stable with respect to variations in the different integration parameters. We were able to show that this is probably related to the greater stability of the intermolecular interaction regions associated with the NCI-ELMO calculations compared to those obtained with the promolecular-NCI strategy.

Afterwards, the new NCI technique has been used in combination with QM/ELMO calculation to evaluate different interactions between ligands and surrounding residues in two protein-ligand complexes. The obtained integral values offer the possibility to rank the different interactions among each other. Together with the integration domains and hydrogen-acceptor distances, we were able to rationalize our findings for the NCI integral values. This application showed that the combination of qualitative and quantitative NCI-ELMO analyses can provide valuable information to identify, classify and rank different non-covalent interactions.

However, further testing will be certainly necessary. For example the integration domains associated with attractive and van der Waals interactions should probably be re-defined, also in light of the results obtained from the NCI-ELMO qualitative analyses. Afterward, we plan to further investigate the protein-ligand complexes with the new NCI-ELMO strategy. As already mentioned, for the PDZ domain, we envisage to also study the interactions of each residue in the ligand with the surrounding protein using the ELMO densities. However, also other protein-ligand complexes or different types of systems could be studied with the new technique. For example, we currently envisage to apply the new strategy to MD simulations, to monitor the evaluation of the interactions along the trajectory and to gain further insights into the dynamics of proteins.

10 The IGM-ELMO technique

10.1 Introduction

Like the NCI index, also the independent gradient model (IGM)[321,415] allows the study of non-covalent interactions. In fact, IGM uses many aspects of the NCI technique, but replaces the reduced density gradient by another descriptor (called δg, see below for more details).

The basis of the IGM is the computation of the molecular electron density and the associated gradient $\nabla\rho$. In a system made up of two fragments (A and B), the gradient is partitioned into individual contributions for each fragment:

$$\frac{\partial\rho}{\partial x} = \left|\frac{\partial\rho_A}{\partial x} + \frac{\partial\rho_B}{\partial x}\right| \tag{10.1}$$

Analogous expressions can be obtained for the other two directions y and z. Based on this expression, the behavior of the gradient in the three dimensional space is evaluated. In the following, this will be described for the example of the diatomic molecule H_2. Figure 10.1 schematically shows the individual atomic densities (ρ_A and ρ_B) and the sum of the two (ρ_{tot}). Starting on the left side of hydrogen A, the density for this atom is increasing rapidly, while the density tail of hydrogen B is also increasing but much less. Nevertheless, in the gray area on the left side of hydrogen A, both gradients of the two atomic electron densities have the same sign. At the nuclear position of hydrogen A, the atomic density of this atom reaches its maximum, afterwards it only decreases. Hence, at this point, the gradient associated with the atomic density of atom A changes its sign, while the atomic density associated with hydrogen B still increases. Therefore, in the white area of Figure 10.1, the two gradients have opposite signs (a feature called contragradience in IGM), leading to an attenuation of the total gradient. Note that in the gray area on the right side of hydrogen B, both atomic densities decrease and the gradients have the same sign. Therefore, no contragradience is associated with the two gray regions in Figure 10.1.

The underlying idea of IGM is to define a corresponding virtual non-interacting reference system, which has the same electron density as the true molecular system, but the interactions between the two fragments A and B are neglected. Mathematically, this corresponds to adding the absolute values of the gradients for the two fragments:

$$\frac{\partial\rho}{\partial x}^{IGM} = \left|\frac{\partial\rho_A}{\partial x}\right| + \left|\frac{\partial\rho_B}{\partial x}\right| \tag{10.2}$$

Equation (10.2) gives the gradient of the non-interacting reference system, which is also called "independent gradient model". Since the absolute values of the gradients are added, the gradient attenuation is canceled. Considering that the total gradient is the sum of all partial derivatives in the directions x, y and z, the norm $\left|\boldsymbol{\nabla}\rho^{IGM}\right|$ resulting from Equation (10.2) is an upper limit to the norm $|\boldsymbol{\nabla}\rho|$ of the actual electron density gradient (Equation (10.1)).

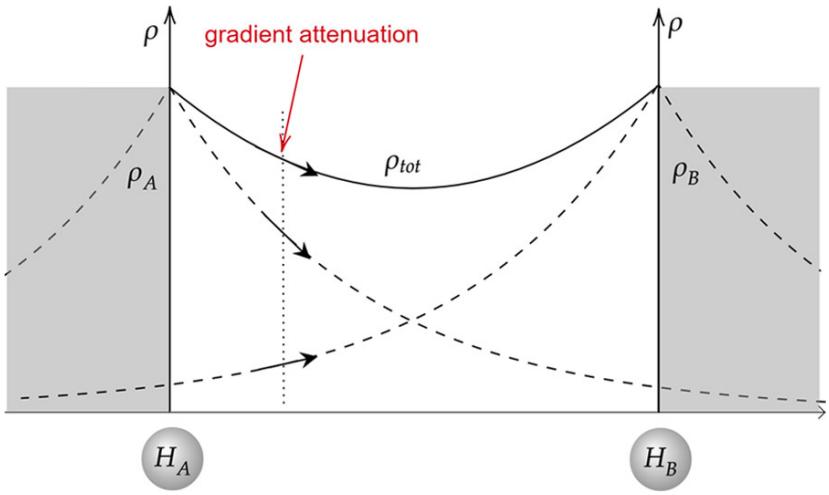


Figure 10.1: Schematic representation of the contragradience between two atomic densities in the H_2 molecule. Reprinted with permission from reference [455]. Copyright 2020 American Chemical Society.

Therefore, the difference between the two

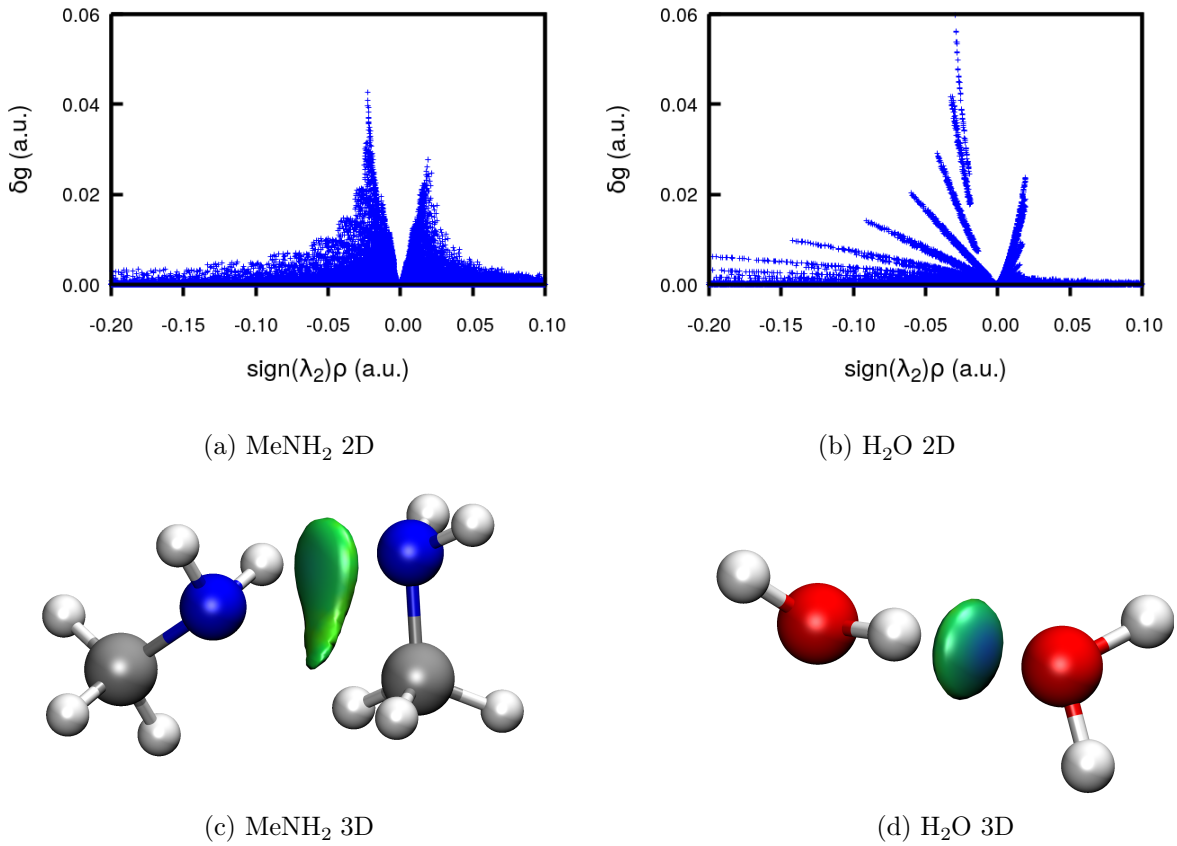
$$\delta \mathrm{g} = \left|\boldsymbol{\nabla}\rho^{IGM}\right| - |\boldsymbol{\nabla}\rho| \tag{10.3}$$

defines the δg descriptor, which allows the accurate identification of interaction regions. Coming back to the example of H_2, as already mentioned, in the gray areas of Figure 10.1, no contragradience is observed. Therefore, in these regions, $\delta g = 0$. In contrast, in the region of gradient attenuation, where the contragradience is observed, $\delta g > 0$. Therefore, positive values of δg exclusively correspond to situations in which interactions occur.

Both the electron density and δg are computed on three dimensional grids. Plotting δg against the "signed" electron density (multiplied again with the sign of the second eigenvalue λ_2 of the electron density Hessian, compare Equation (8.4)), yields 2D IGM plots, which reveal different types of interactions. In Figures 10.2a and 10.2b, the 2D IGM plots for the intermolecular interactions in the methylamine and water dimers are shown, respectively. Note that all the points in the 2D plots exclusively correspond to the non-covalent interaction, in contrast to the corresponding 2D NCI plots (Figure 9.1), where peaks also correspond to covalent interactions. As in the NCI approach, isosurfaces of δg can be drawn and are typically colored according to the sign of λ_2 (attractive interactions in blue, weak ones in green, steric interactions in red). Figures 10.2c and 10.2d show the δg isosurfaces corresponding to the intermolecular interactions in the methylamine and water dimers, respectively.

10.1.1 Quantitative IGM analyses

Furthermore, the IGM can also roughly evaluate the strength associated with an interaction. To this end, the clear definition of the intermolecular interaction regions in IGM is really advantageous. In contrast to NCI, where the RDG displays different interactions simultaneously, the δg descriptor in IGM contains only information about the interaction between the selected fragments A and B. Therefore, in IGM, the 2D IGM plots contain only points associ-



(a) $MeNH_2$ 2D

(b) H_2O 2D

(c) $MeNH_2$ 3D

(d) H_2O 3D

Figure 10.2: 2D and 3D IGM plots for the water and methylamine dimers: Plots of δg against the "signed" electron density for (a) the methylamine dimer and (b) the water dimer and corresponding δg isosurfaces for (c) the methylamine dimer and (d) the water dimer shown with an isovalue of 0.017 a.u.

ated with that interaction (as mentioned above, compare Figure 10.2). Furthermore, none (or little) parametrization is necessary to define the intermolecular interaction region compared to the procedure proposed for the computation of the NCI integrals (see Section 9.1). In fact, the following scheme is applied to compute the IGM integral values:

$$\Delta g = \int_v -\delta g \; dv, \tag{10.4}$$

where the integration is over only those points for which $\lambda_2 < 0$ and $\frac{|\nabla \rho^{IGM}|}{|\nabla \rho|} > 1.2$. The former condition allows us to focus only on attractive non-covalent interactions, while the latter one allows us to focus even more on the interaction since points of very low contragradience are removed.[322]

10.1.2 Types of electron densities in the IGM analyses

Like the NCI index, also the IGM technique crucially depends on the electron density used for the calculations. In its original version, the IGM approach had been only developed for promolecular densities,[415] but shortly after it has been further extended to fully QM densities.[321] However, for the study of large systems, also the IGM approach needed to resort to the crude promolecular density approximation. To allow a more reliable description

of large systems, we have coupled the IGM technique with the ELMO libraries, representing the development ot the new IGM-ELMO approach. Like the NCI-ELMO method, also the IGM-ELMO technique has been thoroughly tested on a number of non-covalent interactions in polypeptides and proteins. The results of this study have been published in reference [322] and will be presented in this chapter.

10.2 Validation on polypeptides

In this section, the results obtained for the validation of the IGM-ELMO technique will be presented and discussed. Similarly to the procedure followed for the qualitative NCI-ELMO analyses (Section 9.2), the IGM-ELMO technique was initially applied to a number of non-covalent interactions in polypeptides and the obtained results were compared to those resulting from promolecular-IGM and IGM-DFT analyses.

10.2.1 Evaluated polypeptide structures

In particular, the following types of interactions and polypeptide structures were considered for the validation of the IGM-ELMO technique:

- A strong hydrogen bond in the high-resolution crystal structure of Leu-enkephalin[314] (86 atoms).
- A T-shaped π-π-stacking in Leu-enkephalin.[314]
- A multiple hydrogen bond between negatively charged asparagine and positively charged arginine residues in the NMR structure of a synthetic peptide that corresponds to an intracellular sequence in the β-adrenergic receptor (fifteen residues, 275 atoms, PDB code: 1DEP).[456]
- Four different π-π-stacking interactions in a chlorinated and a brominated peptide dimer (CSD identifiers: LAKMIX and LAKMOD)[457] and in the corresponding dehalogenated structures (170 atoms).

The first three cases represent non-covalent interactions typically occurring in proteins and were chosen to validate whether the IGM-ELMO technique is able to correctly identify these interactions. Concerning the fourth test case, the peptide dimers were chosen to better assess whether the IGM-ELMO approach is able to correctly quantify the strengths of non-covalent interactions in similar systems. In particular, the final goal was to establish a ranking among the analyzed interactions based on the integrals of the δg values in the "attractive" integration domain (see Equation (10.4)). Therefore, in the following, the results obtained for the peptide dimers will be discussed separately from the interactions in Leu-enkephalin and the synthetic polypeptide 1DEP.

Details about the pre-processing of the structures will be given at the beginning of each section.

10.2.2 Density computation and subsequent IGM analysis

As briefly mentioned above, for all the studied polypeptides, the IGM analyses were based on promolecular, DFT and ELMO densities. The spherical atomic densities used to reconstruct the promolecular densities are stored within the *IGMPLOT* software.[458] Single-point

DFT calculations were performed using the B3LYP functional and all the five basis sets of the ELMO libraries (6-31G, 6-311G, 6-31G(d,p), 6-311G(d,p), and cc-pVDZ). The DFT calculations were carried out with the software *Gaussian16*.[459] Since the structures of Leu-enkephalin, of the synthetic polypeptide 1DEP, and of the dehalogenated peptides contain only standard amino acids, the corresponding ELMO densities were obtained by the transfer of the ELMOs stored in the libraries[105] exploiting all five available basis sets. For the halogenated dimers, ELMOs were computed for the brominated and chlorinated tyrosine residues using the model molecules shown in Appendix D.1 and exploiting again all five basis sets available in the libraries. The resulting ELMOs were added to the database and were transferred to the structures of the halogenated dimers using the *ELMOdb* program.[105] The wavefunction files obtained from the DFT calculations and from the transfers of ELMOs were afterwards passed to the *IGMPLOT* program[458] to perform the IGM analyses. For all the IGM analyses a grid step of 0.1 Å was used.

10.2.3 Non-covalent interactions in Leu-enkephalin and in the synthetic peptide 1DEP

As first test cases for validating the IGM-ELMO technique, three different types of non-covalent interactions were selected. Two of them are present in the crystal structure[314] of Leu-enkephalin. In particular, these are a hydrogen bond between the backbone atoms of residues Tyr1 and Phe4, and a T-shaped π-π interaction between the aromatic side chains of the same residues. The third interaction is a network of multiple hydrogen bonds between the negatively charged residue Asp4 and the positively charged residues Arg1 and Arg11 in the NMR structure[456] of the synthetic peptide 1DEP. Since the high-resolution structure of Leu-enkephalin[314] and the NMR structure[456] of the synthetic polypeptide already contain hydrogen atoms in accurate positions, they were not modified for the subsequent IGM-ELMO calculations.

For these three types of interactions, the corresponding δg-isosurfaces are depicted in panels A, C and E of Figure 10.3. These isosurfaces were obtained from IGM-ELMO calculations (using the cc-pVDZ basis set), and show the expected shapes for these interactions. In particular, for the hydrogen bonds in Leu-enkephalin (Figure 10.3A) and in the synthetic peptide (Figure 10.3C) the corresponding isosurface have the typical lentil shape, while the isosurface corresponding to the T-shaped π-π interaction between the two aromatic rings in Leu-enkephalin is more delocalized. The isosurfaces resulting from B3LYP/cc-pVDZ calculations are hardly distinguishable from the ELMOdb/cc-pVDZ ones in the left side of Figure 10.3 and are therefore not shown. In contrast, some differences can be observed between the promolecular-IGM isosurfaces and the IGM-ELMO and IGM-DFT ones. On the right side of Figure 10.3 (panels B, D and F), the isosurfaces for promolecular-IGM (shown in transluescnet) and IGM-ELMO (shown in opaque) can be compared directly. The promolecular-IGM technique is able to identify all the investigated types of interactions. However, the promolecular δg-isosurfaces are more diffuse than the IGM-ELMO ones. Moreover, for the hydrogen bonds depicted in panels B and F of Figure 10.3, all the promolecular-IGM isosurfaces are clearly shifted towards the acceptor atom of the hydrogen bond.

For all the previously discussed interactions, the corresponding 2D IGM plots are shown

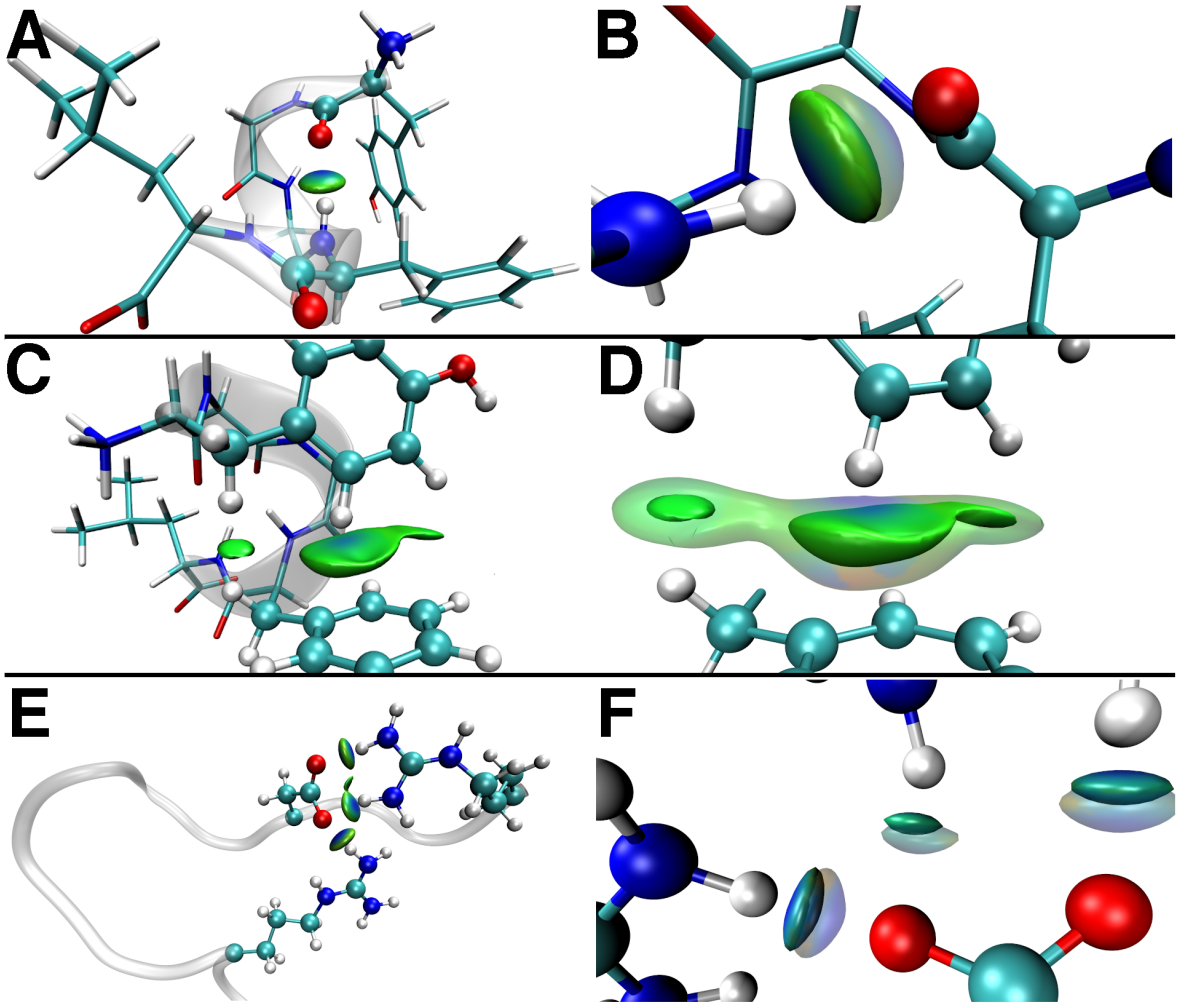


Figure 10.3: δg-isosurfaces associated with (A) the backbone hydrogen bond interaction between residues Tyr1 and Phe4 in Leu-enkephalin at ELMOdb/cc-pVDZ level (zoomed view in (B), with superimposition of the IGM-promolecular isosurface in translucent representation; isovalues set equal to 0.01 a.u.), (C) the T-shaped π-π stacking between residues Tyr1 and Phe4 in Leu-enkephalin at ELMOdb/cc-pVDZ level (zoomed view in (D), with superimposition of the IGM-promolecular isosurface in translucent representation; isovalues set equal to 0.004 a.u.), and (E) the multiple hydrogen bond between the charged residues Asp4, Arg1 and Arg11 in the fifteen-residue polypeptide 1DEP (zoomed view in (F), with superimposition of the IGM-promolecular isosurface in translucent representation; isovalues set equal to 0.015 a.u.). All the isosurfaces are colored according to the BGR scheme over the range -0.05 a.u. $< sign(\lambda_2)\rho <$ 0.05 a.u. Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

in Figure 10.4. In each of the different panels, the plots for promolecular, DFT and ELMO levels (with basis set cc-pVDZ for the QM calculations) are superimposed. Additionally, Figures showing each 2D IGM plot separately for each underlying density and each interaction are given in Appendix D.2. For the hydrogen bond in Leu-enkephalin (panel A and B of Figure 10.4), the ELMO and DFT peaks are very similar to each other, with the ELMO one being slightly shifted to lower absolute values. In contrast, for this interaction, the promolecular peak is clearly shifted to higher absolute values of the "signed" electron density and is also higher in terms of the δg value compared to the DFT peaks.

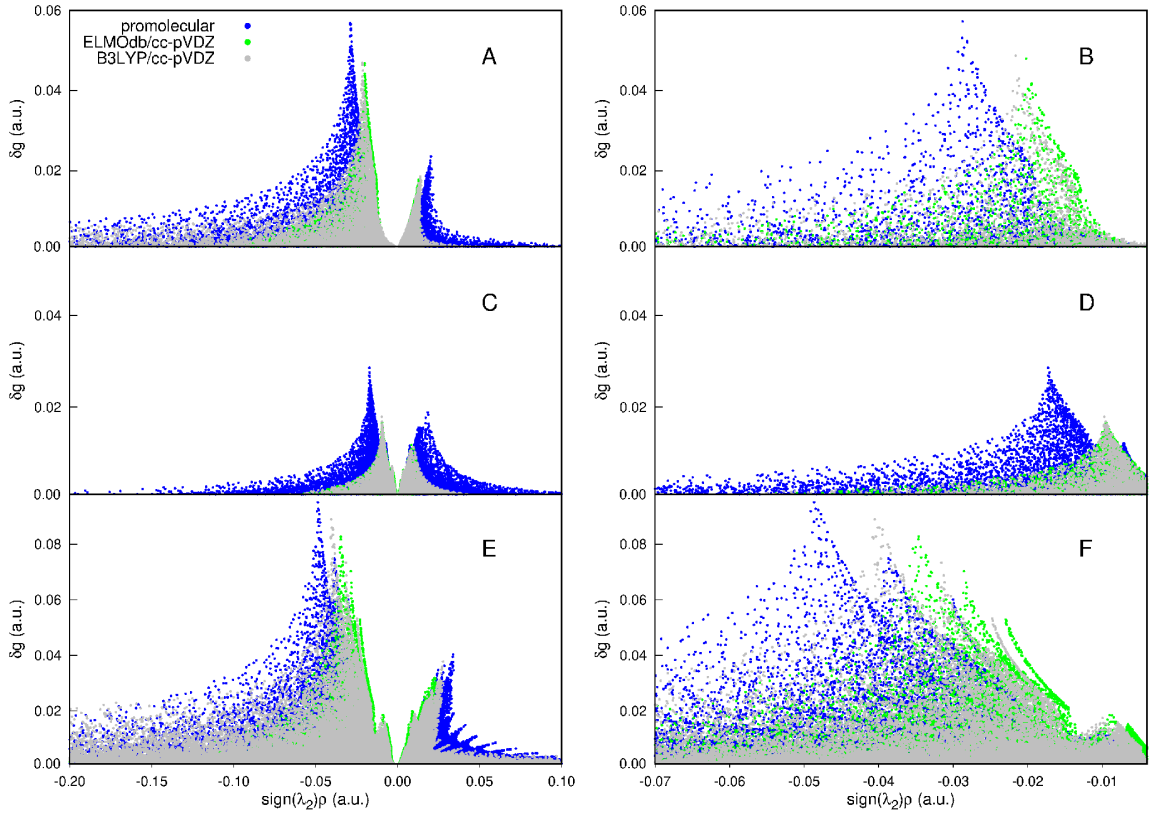


Figure 10.4: Comparison between the 2D IGM plots obtained at promolecular, ELMOdb/cc-pVDZ and B3LYP/cc-pVDZ levels, with zooms on the peaks (in the negative region) in the right panels: (A) and (B) hydrogen bond between residues Tyr1 and Phe4 in Leu-enkephalin, (C) and (D) T-shaped π-π stacking between Tyr1 and Phe4 in Leu-enkephalin, (E) and (F) multiple hydrogen bond between the charged residues Asp4, Arg1 and Arg11 in polypeptide 1DEP. Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

A similar picture is obtained for the π-π interaction in Leu-enkephalin (panel C and D of Figure 10.4). For all the underlying densities, the fingerprint plots have maximum values that are closer to zero in terms of the "signed" electron density and the corresponding peaks are more equal in height. However, the DFT and ELMO peaks are practically indistinguishable, while the promolecular ones are again shifted to higher absolute values of the "signed" electron density and the peaks themselves are higher in terms of δg values.

The fingerprint plots corresponding to the multiple hydrogen bonds in 1DEP are shown in panels E and F of Figure 10.4. In this case, the situation is less clear, since the DFT peaks are located almost precisely in-between the promolecular and the ELMO peaks. However,

comparing also the individual plots shown in Figure D.4, the ELMO peaks seem to be overall more similar to the DFT peaks than the promolecular ones.

To also evaluate the effect of the basis set on the 2D IGM plots, they were superimposed for each basis set and are shown in Figure 10.5. Independently of the considered interactions and the underlying QM levels, the fingerprint plots for the different basis sets are practically identical.

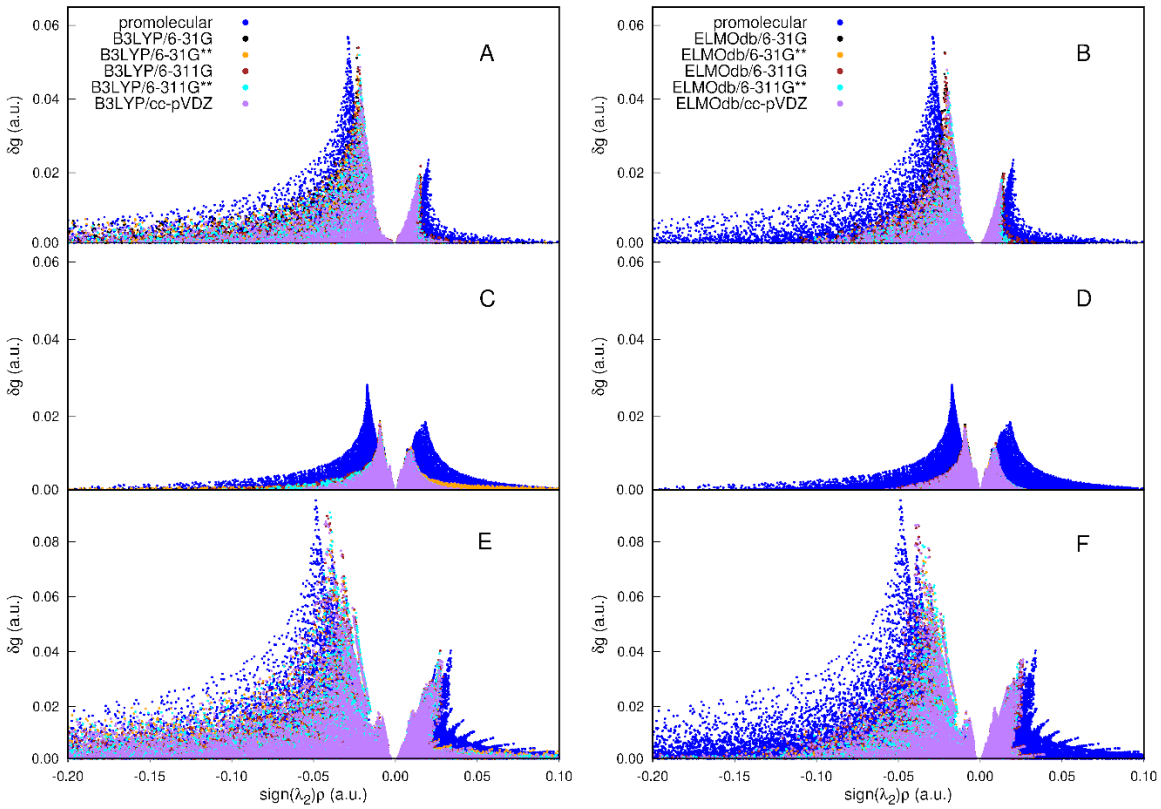


Figure 10.5: Basis set dependence of the 2D IGM plots associated with (A and B) the hydrogen bond between residues Tyr1 and Phe4 in Leu-enkephalin, (C and D) the T-shaped π-π stacking between Tyr1 and Phe4 in Leu-enkephalin, and (E and F) the multiple hydrogen bond between the charged residues Asp4, Arg1 and Arg11 in the polypeptide 1DEP. Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

For a more quantitative analysis of the interactions, integrals of the δg values in the "attractive" integration domain were also computed using Equation (10.4). Henceforth, these integral values will be indicated as Δg values. For the evaluated polypeptides, the obtained Δg values are listed in Table 10.1. For both the interactions in Leu-enkephalin, the promolecular Δg values are significantly larger (in absolute terms) than the DFT ones, indicating that the promolecular approximation overestimates the strengths of these interactions. In contrast, the IGM-ELMO Δg values are significantly closer to the IGM-DFT ones, although the ELMO description slightly underestimates the interaction strengths (as indicated by the smaller (absolute) Δg values). These effects are even more pronounced for the multiple hydrogen bonds in 1DEP, where the promolecular description overestimates and the ELMO one underestimates the interaction strength. Nevertheless, the difference between the ELMO and DFT values is smaller than the one between the promolecular and DFT results. Furthermore, for all three interactions, the different basis sets give rather similar Δg values.

Table 10.1: Δg values (in a.u.) resulting from the IGM analyses on the polypeptides Leu-enkephalin and 1DEP. Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

	Leu-enkephalin		1DEP
	(hydrogen bond)	(π-π interaction)	(hydrogen bonds)
Promolecular	-0.076	-0.180	-0.305
ELMOdb/6-31G	-0.049	-0.071	-0.203
ELMOdb/6-311G	-0.047	-0.075	-0.194
ELMOdb/6-31G(d,p)	-0.050	-0.073	-0.202
ELMOdb/6-311G(d,p)	-0.046	-0.078	-0.188
ELMOdb/cc-pVDZ	-0.052	-0.078	-0.205
B3LYP/6-31G	-0.055	-0.082	-0.238
B3LYP/6-311G	-0.051	-0.084	-0.219
B3LYP/6-31G(d,p)	-0.057	-0.084	-0.249
B3LYP/6-311G(d,p)	-0.053	-0.087	-0.229
B3LYP/cc-pVDZ	-0.056	-0.082	-0.247

Table 10.2: CPU times (in seconds) associated with each step of the IGM analyses of the hydrogen bond and T-shaped π-π interaction in Leu-enkephalin and of the multiple hydrogen bonds in the synthetic polypeptide 1DEP. For each interaction, the first column reports the CPU times required to compute the wavefunctions with *ELMOdb* or *Gaussian16*, while the second column reports the CPU times associated with the IGM analyses that were preformed with *IGMPlot*. In the case of the promolecular approximation, only one value is reported because the calculation of the promolecular electron density and the subsequent IGM analysis are both performed within the *IGMPlot* program. Computations with *Gaussian16* and *IGMPlot* were carried out in parallel using 28 cores. All the calculations were performed on Intel Xeon Gold 6132 2.6 GHz processors of the ROMEO High Performance Computing center (`http://romeo.univ-reims.fr`).

	Leu-enkephalin				1DEP	
	(hydrogen bond)		(π-π interaction)		(hydrogen bonds)	
Promolecular		< 1		2		1
ELMOdb/6-31G	< 1	3	< 1	16	4	26
ELMOdb/6-311G	1	3	1	22	7	36
ELMOdb/6-31G(d,p)	1	4	1	24	9	39
ELMOdb/6-311G(d,p)	1	4	1	27	14	44
ELMOdb/cc-pVDZ	1	4	1	28	10	47
B3LYP/6-31G	115	4	115	20	479	34
B3LYP/6-311G	221	4	221	23	954	40
B3LYP/6-31G(d,p)	244	4	244	25	1141	43
B3LYP/6-311G(d,p)	410	4	410	29	2147	51
B3LYP/cc-pVDZ	394	3	394	31	1836	53

In summary, the analysis of the IGM integral values practically confirm on a quantitative basis the trends observed qualitatively for the 2D IGM plots. In fact, the IGM-ELMO technique gives results that are closer to the IGM-DFT reference than the promolecular-IGM strategy. However, as can be seen in Table 10.2, the IGM-ELMO calculations are significantly faster than the IGM-DFT ones. In fact, the latter already take some minutes for the relatively

small polypeptides considered for these validation tests, whereas the IGM-ELMO analyses always take less than a minute and mostly only a few seconds. The promolecular-IGM is practically instantaneous. However, also the timings associated with the *ELMOdb* program could be improved significantly in the future through the implementation of a parallelized version of the software. Furthermore, from Table 10.2, it appears that the time required for the IGM analyses with *IGMPlot* is practically independent of the underlying QM calculations. This shows that the bottleneck for IGM analyses is clearly the computation of the wavefunction, for which the computational cost will certainly increase when larger systems (such as proteins) are studied. Therefore, there is clearly a need for a technique like IGM-ELMO that represents a good compromise between accuracy and computational cost. In fact, such analyses of non-covalent interactions in proteins will be presented in Section 10.3. However, before discussing the application of the technique to such large systems, its capability to rank the non-covalent interactions in similar systems will be described and commented.

10.2.4 Non-covalent interactions in halogenated peptide dimers

In this second part of the validation, π-π stacking interactions in halogenated and dehalogenated peptide dimers will be discussed. The halogenated peptides were extracted from two crystal structures[457] of the halogenated sequence 7-12 (DSGYEV) of the flexible amyloid-β peptide. In these crystal structures, the tyrosine residue in position Y10 is either chlorinated or brominated at the *ortho*-positions with respect to the hydroxyl group in the aromatic ring. These halogenated tyrosine residues form π-π stacking interactions with the halogenated tyrosine residue in the neighboring chain. Therefore, dimers of the brominated peptide (DSGY(Br)EV) and of the chlorinated peptide (DSGY(Cl)EV) were extracted from the crystal structures and are shown in Figures 10.6A and 10.7A.

Since both X-ray crystal structures were refined using the IAM (which provides systematically too short element-hydrogen bond lengths, for details about this refinement model see Chapter 2), the hydrogen atoms were removed from these structures and afterwards added with the software *MolProbity*,[460–462] which adds hydrogen atoms at neutron distances to their bonding partners.

Furthermore, structures of dehalogenated dimers were also prepared, where the halogen atoms were substituted with hydrogen atoms. In practice, this was done by removing the halogen atoms from the structures of the halogenated dimers and afterwards adding the hydrogen atoms with the software *MolProbity*. The resulting structures are shown in Figures 10.6D and 10.7D. It is worth noting that the geometries of the two dehalogenated dimers are different because they derive from the two different crystal structures of the halogenated polypeptides. Therefore, to avoid a bias due to different geometries, the results of the IGM analyses will be compared only between each pair of halogenated and dehalogenated dimer.

The π-π stacking interactions in the four peptide dimers were analyzed with the promolecular-IGM, IGM-DFT and IGM-ELMO strategies, considering again all the five basis sets of the ELMO libraries in the computations at DFT and ELMO levels. The obtained δg isosurfaces are shown in Figure 10.6 for the brominated and debrominated peptide dimers and in Figure 10.7 for the chlorinated and dechlorinated ones. For all four dimers, the IGM analysis is able to identify the π-π stacking interactions, which correspond to the large green

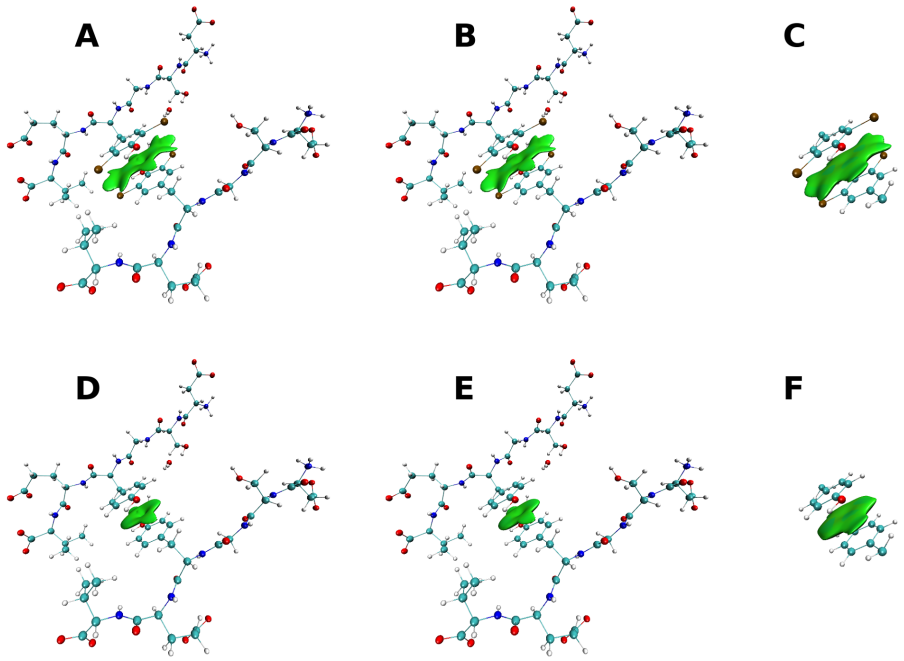


Figure 10.6: δg-isosurfaces associated with the π-π interactions in the dimers of the (A-C) brominated and (D-F) debrominated polypeptides based on B3LYP/cc-pVDZ (A and D), ELMOdb/cc-pVDZ (B and E) and promolecular (C and F) electron densities. All the isosurfaces correspond to the 0.004 a.u. isovalue and are colored according to the BGR scheme over the range -0.08 a.u. $< sign(\lambda_2)\rho <$ 0.08 a.u. Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

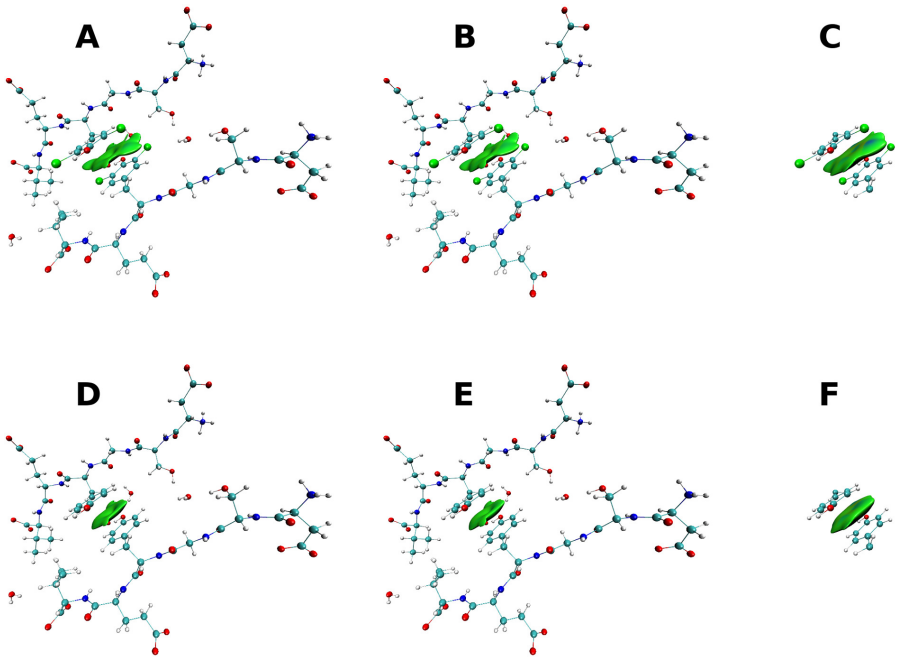


Figure 10.7: δg-isosurfaces associated with the π-π interactions in the dimers of the (A-C) chlorinated and (D-F) dechlorinated polypeptides based on B3LYP/cc-pVDZ (A and D), ELMOdb/cc-pVDZ (B and E) and promolecular (C and F) electron densities. All the isosurfaces correspond to the 0.004 a.u. isovalue and are colored according to the BGR scheme over the range -0.08 a.u. $< sign(\lambda_2)\rho <$ 0.08 a.u.

δg isosurfaces in between the aromatic rings. Irrespective of the underlying density, the IGM approach provides isosurfaces that are more extended for the halogenated dimers than for the dehalogenated ones. However, the IGM-DFT and IGM-ELMO strategies provide isosurfaces that are very similar to each other, whereas the ones obtained with the promolecular-IGM technique are always more diffuse.

The corresponding 2D IGM plots for the π-π stacking interactions are depicted in Figure 10.8, where the peaks obtained at B3LYP/cc-pVDZ, ELMOdb/cc-pVDZ and promolecular levels are superimposed. Individual plots for each method are shown in Appendix D.3. For all the different interactions, the fingerprint plots show two similar peaks close to zero, which are typically obtained for weak interactions such as the π-π stacking. In all four cases, the plots obtained with the IGM-DFT or IGM-ELMO strategies are very similar to each other. This is particularly evident for the chlorinated and dechlorinated dimers, for which the corresponding peaks overlap almost perfectly. In contrast, the promolecular-IGM technique provides peaks that are shifted to higher absolute values of the "signed" electron density and that are higher in terms of the δg values compared to the peaks obtained with the IGM-DFT approach. The same trends can be also observed for the brominated and debrominated dimers. However, in this case the peaks resulting from IGM-ELMO are also slightly higher and marginally shifted to higher absolute values of the "signed" electron density compared to the IGM-DFT peaks. Nevertheless, these differences are much smaller than between the promolecular and the DFT peaks.

As already observed in the validation of the other polypeptides, also in the cases of the halogenated and dehalogenated dimers the influence of the basis sets on the 2D IGM plots is negligible, as can be seen in Figures D.7 and D.8.

Finally, we carried out a more quantitative analysis for the four peptide dimers. To accomplish this task, the Δg values resulting from the integration of the δg peaks in Figure 10.8 are reported in Table 10.3. In general, the Δg values are associated with the strengths of the interactions, and a larger (absolute) value of Δg corresponds to a stronger interaction.

Furthermore, to compute the interaction energies associated with the different π-π stackings, the tyrosine side chains were extracted from the structures of each considered dimer, by cutting each tyrosine residue at the carbon β positions. The valency of these atoms were saturated with hydrogen atoms, for which the positions were optimized while keeping the geometry of the remaining side chains. The interaction energies were subsequently calculated on the optimized reduced geometries exploiting the supermolecular approach with the counterpoise correction (compare Section 8.2.2). More precisely, the energies were computed using the B3LYP functional with Grimme's empirical dispersion D3 correction[395] and Becke-Johnson damping[463] (B3LYP-D3(BJ) level) in combination with all the five basis sets (6-31G, 6-311G, 6-31G(d,p), 6-311G(d,p), and cc-pVDZ) considered also for the QM-based IGM analyses. The resulting interaction energies are reported in the second part of Table 10.3.

From Table 10.3 it can be observed that the Δg values and the interaction energies associated with the π-π stackings follow the same trends. In fact, both quantities indicate that the π-π interactions are stronger in the halogenated dimers than in the dehalogenated ones. This is consistent with the more extended δg isosurfaces observed for the halogenated dimers in Figures 10.6 and 10.7. Moreover, both the Δg values and the interaction energies are showing

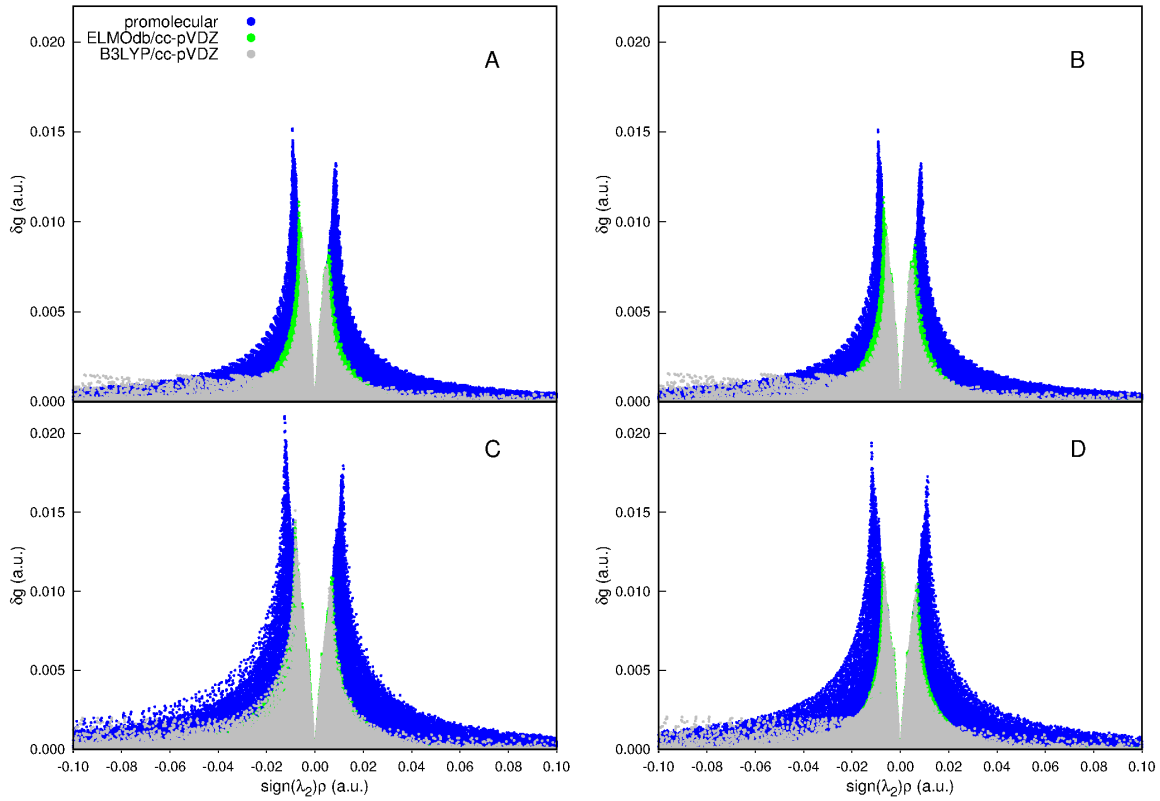


Figure 10.8: Comparison between the 2D IGM plots obtained at promolecular, ELMOdb/cc-pVDZ and B3LYP/cc-pVDZ levels for the π-π interactions in the dimers of the (A) brominated, (B) debrominated, (C) chlorinated, and (D) dechlorinated polypeptides. Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

that the π-π stacking interaction is stronger in the brominated dimer than in the chlorinated one.

For all the four interactions, the Δg values based on ELMO and DFT densities are very similar. In contrast, the absolute Δg values obtained from the promolecular-IGM calculations are significantly larger than those resulting from the QM-based IGM analyses. These observations are consistent with the differences observed in the 2D IGM plots.

In summary, all the validation tests have shown a high similarity between the results obtained with the IGM-ELMO and IGM-DFT approaches. Furthermore, the calculations on the halogenated and dehalogenated dimers confirmed that also the IGM-ELMO technique is able to correctly evaluate and rank the strength of interactions in chemically similar systems.

Table 10.3: Δg values (in a.u.) and interaction energies (in kcal/mol) for the π-π interactions between tyrosine residues of the halogenated and dehalogenated peptide dimers. Adapted with permission from reference [322]. Copyright 2021 American Chemical Society.

Δg values (in a.u.)				
	Brominated dimer	Debrominated dimer	Chlorinated dimer	Dechlorinated dimer
Promolecular	-0.340	-0.255	-0.331	-0.281
ELMOdb/6-31G	-0.194	-0.126	-0.179	-0.140
ELMOdb/6-311G	-0.194	-0.125	-0.181	-0.142
ELMOdb/6-31G(d,p)	-0.194	-0.126	-0.174	-0.139
ELMOdb/6-311G(d,p)	-0.194	-0.127	-0.178	-0.144
ELMOdb/cc-pVDZ	-0.206	-0.132	-0.190	-0.150
B3LYP/6-31G	-0.205	-0.127	-0.180	-0.135
B3LYP/6-311G	-0.204	-0.126	-0.180	-0.134
B3LYP/6-31G(d,p)	-0.207	-0.130	-0.176	-0.139
B3LYP/6-311G(d,p)	-0.205	-0.128	-0.175	-0.137
B3LYP/cc-pVDZ	-0.212	-0.125	-0.178	-0.133

Interaction energies (in kcal/mol)				
	Brominated dimer	Debrominated dimer	Chlorinated dimer	Dechlorinated dimer
B3LYP-D3(BJ)/6-31G	-11.28	-3.81	-8.73	-3.69
B3LYP-D3(BJ)/6-311G	-11.74	-4.28	-9.53	-4.32
B3LYP-D3(BJ)/6-31G(d,p)	-11.15	-3.70	-8.50	-3.60
B3LYP-D3(BJ)/6-311G(d,p)	-11.70	-4.22	-9.33	-4.28
B3LYP-D3(BJ)/cc-pVDZ	-11.30	-4.09	-8.86	-4.13

10.3 Application to proteins

Following the successful validation of the IGM-ELMO technique on the polypeptides, the new method was applied to study different non-covalent interactions in proteins. In particular, the following structures and types of interactions were considered:

- In the crystal structure[436] of human erythrocytic ubiquitin (1231 atoms, PDB code: 1UBQ):
  - a C$-$H$\cdots\pi$ interaction between residues Leu50 and Tyr59.
- In the crystal structure[442] of human carbonic anhydrase II (1477 atoms, PDB code: 3KS3):
  - a hydrogen bond between residues Asn61 and Gly63;
  - a lone pair (n-π^*) interaction between the oxygen atom of the Asn61 sidechain and the carbonyl group of the Asn61 backbone;
  - a multiple hydrogen bond between the charged residues Arg58 and Glu69.

The structures of the proteins were taken from the PDB.[219] Before the calculation of the densities, if necessary, the following modifications were made to the PDB files: (i) only one of the disordered conformations of the proteins were selected; (ii) the correct protonation states

were assigned according to the pH value of crystallization; and (iii) missing hydrogen atoms were added using the *tleap* software in *AMBER*.[148]

Promolecular-IGM and IGM-ELMO (with all five basis sets available in the ELMO libraries) analyses were performed for all four interactions listed above. In the following, the results for each interaction will be discussed separately.

In the first protein (human erythrocytic ubiquitin) we studied the C−H$\cdots\pi$ interaction with the promolecular-IGM and IGM-ELMO techniques. The corresponding δg isosurface resulting from the IGM analysis at *ELMOdb*/cc-pVDZ level is shown in Figure 10.9A. Similar to the δg isosurface observed for the π-π interaction in Leu-enkephalin (compare Figure 10.3A), also the isosurface in Figure 10.9A is rather delocalized. Furthermore, as can be seen from Figure 10.9B, the promolecular-IGM δg isosurface is more diffuse than the IGM-ELMO one, which is again analogous to the results for the π-π interaction in Leu-enkephalin (see Figure 10.3B).

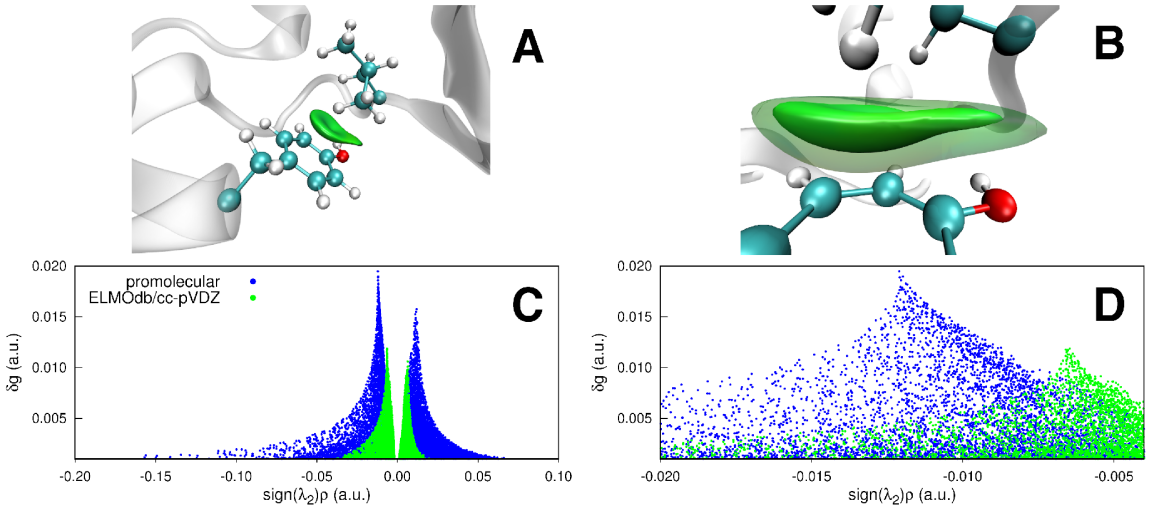


Figure 10.9: C−H$\cdots\pi$ interaction between residues Leu50 and Tyr59 in ubiquitin (PDB ID: 1UBQ): (A) δg = 0.005 a.u. isosurface obtained with the ELMOdb/cc-pVDZ electron density and colored according to the BGR scheme over the range -0.05 a.u. $< sign(\lambda_2)\rho <$ 0.05 a.u.; (B) ELMOdb versus promolecular (translucent) δg = 0.005 a.u. isosurfaces; (C) comparison between the 2D IGM plots obtained at promolecular and ELMOdb/cc-pVDZ levels, with a zoom on the peaks (negative region) in (D). Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

The 2D IGM plots for the C−H$\cdots\pi$ interaction are superimposed in Figure 10.9C and have the typical shape of weak interactions with two peaks that occur at values close to zero. However, the promolecular-IGM peaks are higher in terms of δg value and are shifted toward higher absolute values of the "signed" electron density compared to the IGM-ELMO ones. This is highlighted in Figure 10.9D, which shows a zoom into the attractive (negative) range of the "signed" electron density. Also these plots are very similar to the ones for the π-π interaction in Leu-enkephalin.

Finally, in Table 10.4, the integrated Δg values are reported for the C−H$\cdots\pi$ interaction. Independently of the basis sets used in the computation of the ELMO densities, all the IGM-ELMO analyses provided Δg values that are very similar and significantly smaller (in absolute terms) than the one obtained with the promolecular-IGM strategy. Also these trends

Table 10.4: Δg values (in a.u.) associated with the non-covalent interactions in proteins. Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

	1UBQ	3KS3		
	C−H$\cdots\pi$ interaction	Local hydrogen bond	n-π^* interaction	Multiple hydrogen bond
Promolecular	-0.137	-0.080	-0.103	-0.194
ELMOdb/6-31G	-0.063	-0.053	-0.042	-0.125
ELMOdb/6-311G	-0.064	-0.050	-0.042	-0.117
ELMOdb/6-31G(d,p)	-0.064	-0.054	-0.041	-0.127
ELMOdb/6-311G(d,p)	-0.067	-0.051	-0.042	-0.115
ELMOdb/cc-pVDZ	-0.068	-0.056	-0.040	-0.127

were observed before for the π-π interaction in Leu-enkephalin.

In the second protein (human carbonic anhydrase II) we have first studied a local hydrogen bond. The associated δg isosurface obtained at ELMOdb/cc-pVDZ level is depicted in Figure 10.10A and has the typical disk shape usually observed for hydrogen bonds. Figure 10.10B shows the difference between the IGM-ELMO isosurface and the promolecular-IGM one, where the promolecular-IGM isosurface is shifted slightly towards the acceptor atom of the hydrogen bond. Furthermore, the corresponding 2D IGM plots are shown in panels C and D of Figure 10.10. The promolecular-IGM peaks are higher and shifted to greater absolute values of the "signed" electron density. All these observations are fully consistent and very similar to the findings for the hydrogen bond in Leu-enkephalin (compare panels A and B in Figures 10.3 and 10.4), which highlights the similarity between the two hydrogen bonds. This is further confirmed on a more quantitative basis by the IGM integral values, which are reported in Table 10.4. In fact, also these values are very similar to the ones obtained for the hydrogen bond in Leu-enkephalin (see Table 10.1). The values obtained from the promolecular approximation are again significantly larger than those resulting from IGM analyses based on ELMO densities, which is observed independently of the adopted basis set.

The local hydrogen bond in human carbonic anhydrase II is accompanied by an n-π^* interaction that is formed between the carbonyl carbon lone pair in the side chain of Asn61 and the oxygen atom in the backbone of the same residue. In fact, the latter is also the acceptor atom of the previously discussed hydrogen bond. For more general details about the local hydrogen bond and the n-π^* interaction, see Section 9.2.3.3, where the same interaction is analyzed with the NCI method. The results obtained with the IGM technique are reported in Figure 10.11. Interestingly, the δg isosurface obtained at ELMO/cc-pVDZ level has a disk shape similar to the one for a hydrogen bond (see panel A). In contrast, as can be seen in panel B, the promolecular-IGM analysis provides a larger and differently oriented δg isosurface that is also shifted towards the backbone oxygen atom of Asn61. Nevertheless, the corresponding 2D IGM plots (Figure 10.11C and D) show peaks that are again rather similar to the fingerprint plots usually obtained for hydrogen bonds. Concerning the difference between the peaks obtained at promolecular and ELMO levels, the latter are again lower and shifted towards lower absolute values of the "signed" electron density. Interestingly, the difference between the δg isosurfaces obtained at promolecular and ELMO levels is mirrored

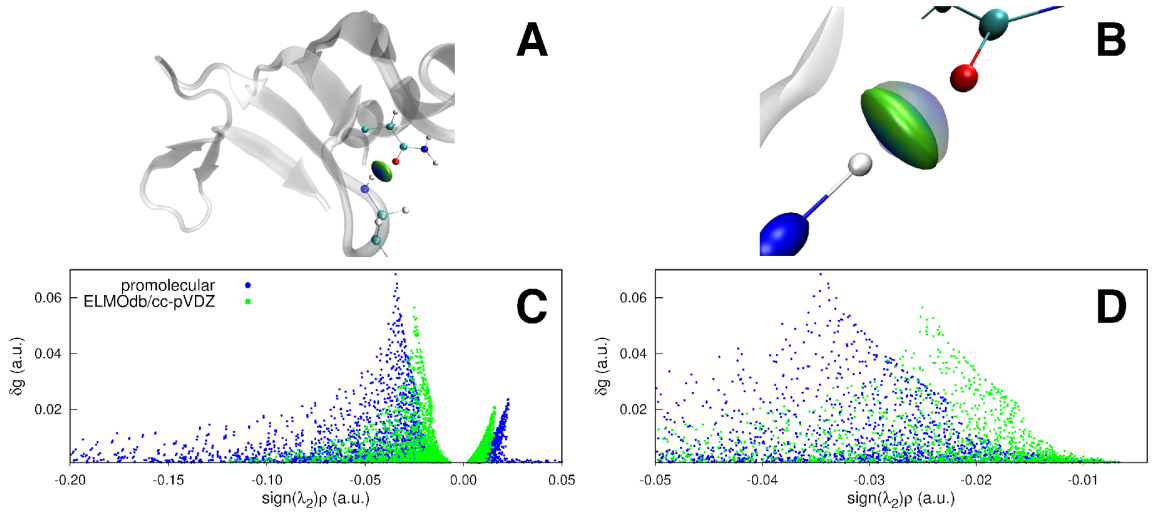


Figure 10.10: Local hydrogen-bond between residues Asn61 and Gly63 in human carbonic anhydrase II (PDB ID: 3KS3): (A) δg = 0.01 a.u. isosurface obtained with the ELMOdb/cc-pVDZ electron density and colored according to the BGR scheme over the range -0.05 a.u. $< sign(\lambda_2)\rho <$ 0.05 a.u.; (B) ELMOdb versus promolecular (translucent) δg = 0.01 a.u. isosurfaces; (C) comparison between the 2D IGM plots obtained at promolecular and ELMOdb/cc-pVDZ levels, with a zoom on the peaks (negative region) in (D). Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

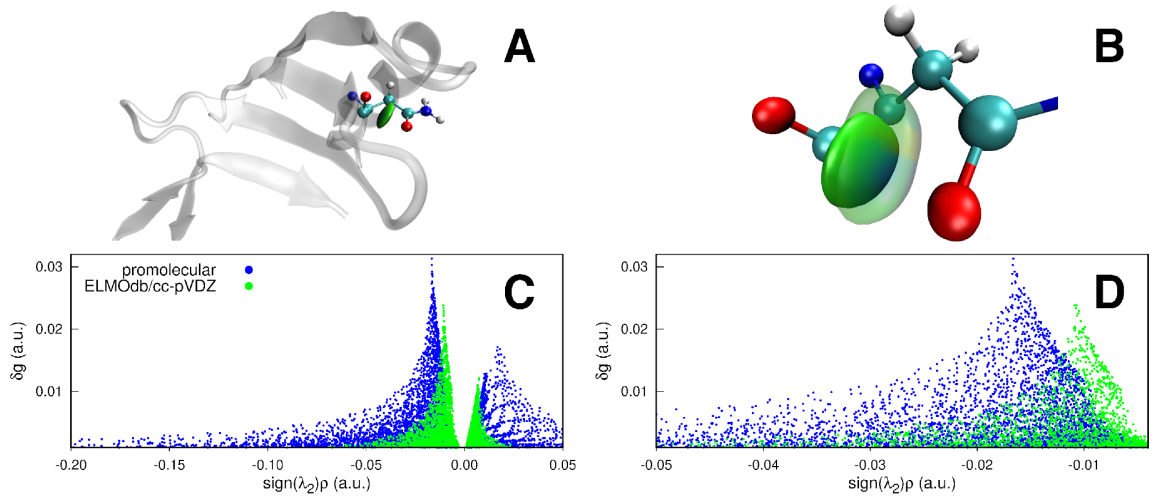


Figure 10.11: n-π^* interaction between the lone pair on the oxygen atom of the Asn61 sidechain and the π^* molecular orbital associated with the carbonyl group of the Asn61 backbone in human carbonic anhydrase II (PDB ID: 3KS3): (A) δg = 0.005 a.u. isosurface obtained with the ELMOdb/cc-pVDZ electron density and colored according to the BGR scheme over the range -0.05 a.u. $< sign(\lambda_2)\rho <$ 0.05 a.u.; (B) ELMOdb versus promolecular (translucent) δg = 0.005 a.u. isosurfaces; (C) comparison between the 2D IGM plots obtained at promolecular and ELMOdb/cc-pVDZ levels, with a zoom on the peaks (negative region) in (D). Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

in the Δg values reported in Table 10.4. In fact, the promolecular Δg values are (in absolute values) significantly larger than the ELMO ones (by more than a factor of two), and are

even larger than the promolecular Δg value obtained for the local hydrogen bond. Hence, the promolecular approximation suggests that the n-π^* interaction is stronger than the local hydrogen bond, which is in contradiction with the results obtained at ELMO level and with the trends described in the literature. In fact, Bartlett *et al.* report that the n-π^* interaction makes up only approximately from 5 to 25% of the interaction energy of a hydrogen bond.[380]

As a last example of a non-covalent interaction in proteins, a multiple hydrogen bond interaction in human carbonic anhydrase II is analyzed with the IGM-ELMO and promolecular-IGM techniques. The corresponding δg isosurfaces are reported in panels A and B of Figure 10.12. As in the other cases, the promolecular isosurfaces are more diffuse and shifted toward the acceptor atoms of the hydrogen bonds when compared to the ELMO ones. Also the 2D IGM plots show the same trends observed before, with higher promolecular peaks occurring at larger absolute values of the "signed" electron density. Finally, also the absolute Δg value (see Table 10.4) resulting from the promolecular-IGM analysis is larger than all the values obtained at ELMO levels, irrespective of the adopted basis set. All these observations are consistent with the results obtained for the multiple hydrogen bond interaction in the synthetic peptide 1DEP (compare panels E and F in Figures 10.3 and 10.4).

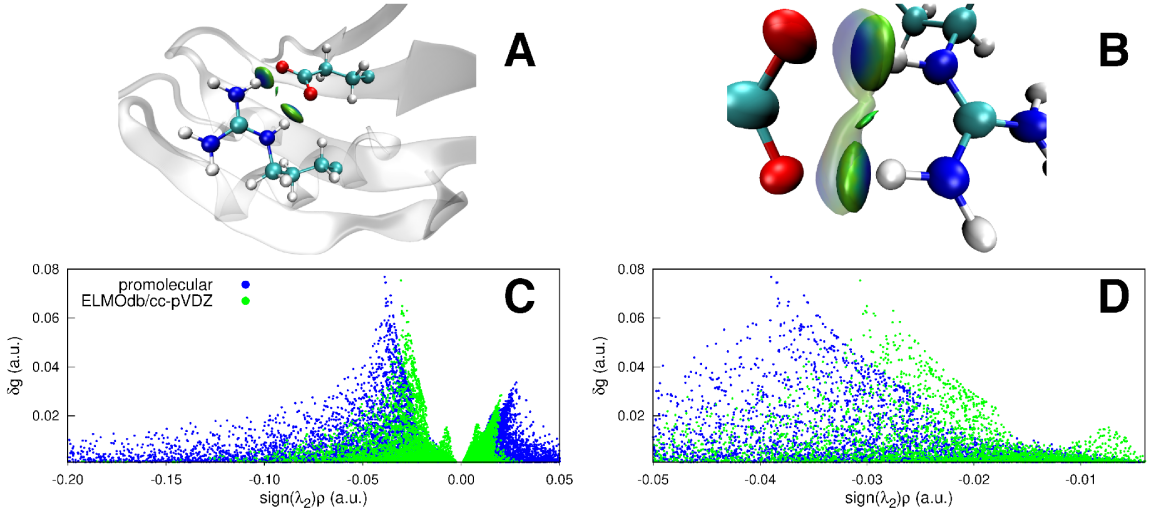


Figure 10.12: Multiple hydrogen bond interaction between residues Arg58 and Glu69 in human carbonic anhydrase II (PDB ID: 3KS3): (A) δg = 0.014 a.u. isosurface obtained with the ELMOdb/cc-pVDZ electron density and colored according to the BGR scheme over the range -0.05 a.u.< $sign(\lambda_2)\rho$ < 0.05 a.u.; (B) ELMOdb versus promolecular (translucent) δg = 0.014 a.u. isosurfaces; (C) comparison between the 2D IGM plots obtained at promolecular and ELMOdb/cc-pVDZ levels, with a zoom on the peaks (negative region) in (D). Reprinted with permission from reference [322]. Copyright 2021 American Chemical Society.

In summary, for all the interactions analyzed in the considered proteins, both the promolecular-IGM and IGM-ELMO techniques provide results that are very similar to the ones obtained for the polypeptides. Therefore, one could speculate that, if fully QM calculations at DFT level were feasible for the studied proteins, they should provide results similar to those obtained at ELMO level.

10.4 Conclusions

In this chapter, the new IGM-ELMO technique was described. The underlying coupling of the IGM method with the ELMO libraries allows the use of quantum mechanically rigorous electron densities of macromolecules, instead of the crude promolecular approximation.

The new technique was validated on a set of non-covalent interactions in small polypeptides and peptide dimers. The results obtained with the IGM-ELMO strategy were compared to those resulting from promolecular-IGM and IGM-DFT analyses. The validation tests showed that the novel IGM-ELMO method provides results that are very close to the ones obtained at DFT level, whereas the promolecular approach gives results that are markedly different. As the qualitative and quantitative analyses show, the promolecular strategy constantly overestimates the interaction signatures and the strengths of the interactions. Moreover, the validation tests also indicated that the new IGM-ELMO approach is capable of correctly assessing relative strengths of non-covalent interactions in chemically similar systems.

Furthermore, better results can be obtained with the IGM-ELMO technique at only a slightly larger computational cost compared to the promolecular-IGM approach, which allowed us to apply the IGM-ELMO technique to study typical non-covalent interactions in macromolecules. The performed analyses provided results that were fully consistent with those obtained from the validation tests on smaller peptides. In fact, for interactions of similar types, completely analogous trends were observed.

In summary, the new IGM-ELMO technique has proven to be a reliable tool for identifying and quantifying non-covalent interactions in polypeptides and proteins.

11 Summary and conclusions of Part II

Non-covalent interactions define the structures of small and large molecules in the liquid and solid phase and play a key role in the processes of life. Therefore, many different strategies have been developed to study these unique interactions by means of quantum chemistry. Some of these techniques have been described in Chapter 8. Methods of particular interest are those that allow us to identify non-covalent interactions in real space, classify their various types and, ideally, also quantify their strengths. Two approaches that aim at accomplishing all these three tasks, are the NCI and IGM techniques. Both of them are based on the analysis of the electron density and of a electron density gradient-based quantity.

To compute the electron densities of large systems such as proteins, both techniques had to resort to the promolecular approximation. To allow for a more reliable analysis also for these large systems, we have coupled the NCI and IGM techniques with the ELMO libraries.[320,322] The capabilities of the resulting NCI-ELMO and IGM-ELMO approaches have been shown and analyzed in this part of the thesis.

In particular, in the first part of Chapter 9, it has been shown that the NCI-ELMO technique is able to provide results that are very close to those of NCI-DFT calculations but at a significantly reduced computational cost. Moreover, for all the studied interactions, the NCI-ELMO results are more reliable than the promolecular ones. This allowed us to apply the NCI-ELMO technique to a variety of non-covalent interactions in proteins, which were all successfully detected and classified.

Furthermore, to provide a more quantitative NCI analysis, in the second part of Chapter 9, the parameters necessary to compute NCI integrals have been (re-)evaluated for promolecular and ELMO densities. Also this parametrization showed that ELMOs are the tools of choice for assessing the strengths of non-covalent interactions with the NCI technique. Based on these encouraging results, we considered two examples of protein-ligand complexes and we estimated the strengths of interactions established by different protein residues with the ligands.

Likewise, also the IGM-ELMO technique is able to reliably detect non-covalent interactions in biosystems, as it has been demonstrated in Chapter 10 for different small polypeptides and proteins. Moreover, it has been shown that the IGM-ELMO technique also allows the ranking of interactions in similar systems according to their strengths. In analogy with the NCI-ELMO technique, also the IGM-ELMO method provides results that are in good agreement with analyses based on DFT calculations, but at a significantly reduced computational cost. Therefore, also the IGM-ELMO technique can be applied to study non-covalent interactions in proteins, as shown by the examples of four non-covalent interactions that are typically found in these macromolecules.

In conclusions, both the NCI-ELMO and IGM-ELMO techniques allow the identification and classification of non-covalent interactions in polypeptides and proteins, and also the possibility of ranking the interactions according to their strengths.

For this reason, we envisage that in the future both techniques could provide useful descrip-

tors to be exploited in molecular docking calculations or virtual high-throughput screenings for rational drug design. Furthermore, since protein structures are intrinsically dynamic, both methods could be also applied to follow the appearance and disappearance of non-covalent interactions along MD trajectories.

Appendix

A Appendix to Chapter 4

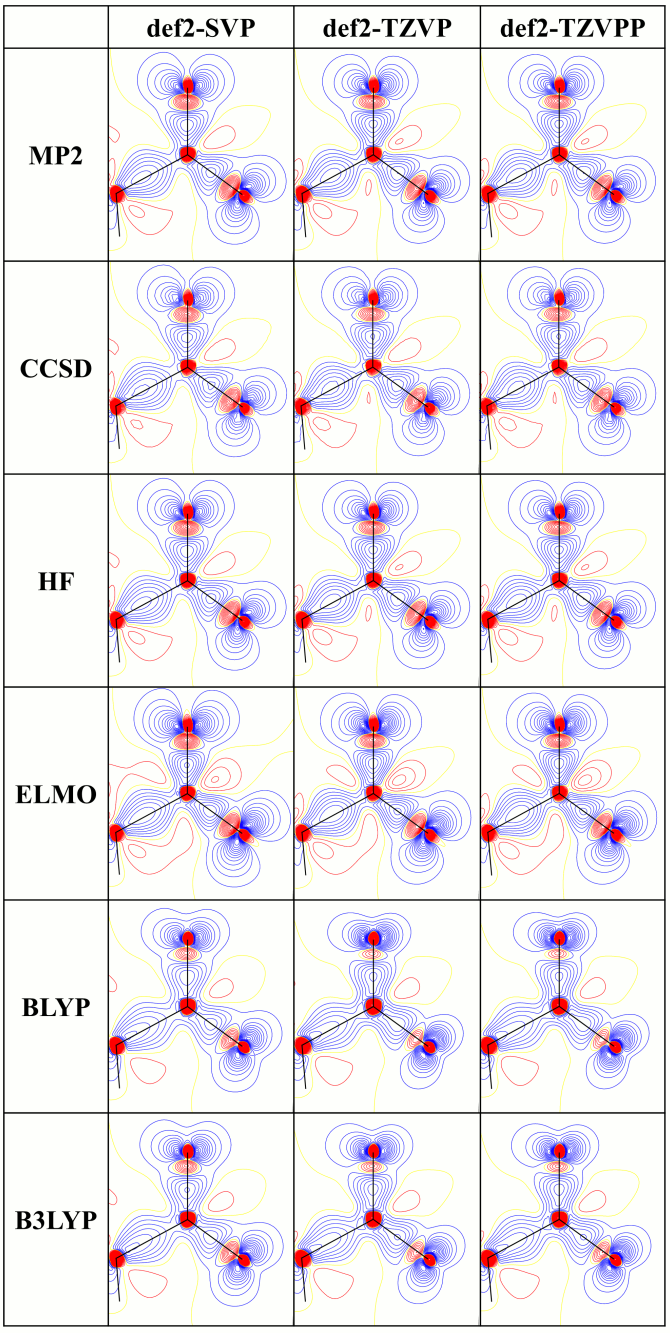


Figure A.1: HAR deformation densities in the plane of the carboxylate group of *L*-alanine. Contour level: 0.01 eÅ^{-3}; blue = positive; red = negative.

Table A.1: Statistical analysis of the C−H bond lengths obtained from the different Hirshfeld atom refinements: (i) average bond lengths $\langle \boldsymbol{r}_{\mathrm{HAR}}(\mathrm{C{-}H})\rangle$, (ii) mean ratios $\langle \boldsymbol{r}_{\mathrm{HAR}}/\boldsymbol{r}_{\mathrm{neutron}}\rangle$ and (iii) the mean absolute differences $\langle|\boldsymbol{r}_{\mathrm{HAR}} - \boldsymbol{r}_{\mathrm{neutron}}|\rangle$. For each basis set, the first column refers to the value of the quantity and the second column to the corresponding standard deviation upon averaging. The average neutron C−H bond length is 1.095 Å with a standard deviation of 0.003 Å.

$\langle \boldsymbol{r}_{\mathrm{HAR}}(\mathrm{C{-}H})\rangle$/Å						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	1.090	0.002	1.095	0.006	1.097	0.005
CCSD	1.088	0.002	1.093	0.006	1.095	0.006
HF	1.091	0.005	1.094	0.009	1.095	0.009
ELMO	1.090	0.007	1.094	0.012	1.095	0.013
BLYP	1.089	0.003	1.091	0.006	1.093	0.005
B3LYP	1.091	0.002	1.093	0.006	1.094	0.005

$\langle \boldsymbol{r}_{\mathrm{HAR}}/\boldsymbol{r}_{\mathrm{neutron}}\rangle$						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.996	0.005	1.000	0.007	1.001	0.007
CCSD	0.994	0.004	0.998	0.008	1.000	0.007
HF	0.996	0.005	0.998	0.009	1.000	0.009
ELMO	0.995	0.006	0.998	0.011	1.000	0.012
BLYP	0.995	0.006	0.996	0.008	0.998	0.007
B3LYP	0.996	0.004	0.998	0.007	0.999	0.007

$\langle\lvert\boldsymbol{r}_{\mathrm{HAR}} - \boldsymbol{r}_{\mathrm{neutron}}\rvert\rangle$/Å						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.005	0.005	0.007	0.002	0.006	0.003
CCSD	0.007	0.005	0.007	0.003	0.007	0.002
HF	0.004	0.005	0.007	0.006	0.007	0.006
ELMO	0.005	0.007	0.009	0.007	0.010	0.008
BLYP	0.006	0.006	0.008	0.004	0.007	0.004
B3LYP	0.005	0.005	0.006	0.004	0.006	0.003

Table A.2: Statistical analysis of the N−H bond lengths obtained from the different Hirshfeld atom refinements: (i) average bond lengths $\langle \boldsymbol{r}_{\mathrm{HAR}}(\mathrm{N-H})\rangle$, (ii) mean ratios $\langle \boldsymbol{r}_{\mathrm{HAR}}/\boldsymbol{r}_{\mathrm{neutron}}\rangle$ and (iii) the mean absolute differences $\langle|\boldsymbol{r}_{\mathrm{HAR}} - \boldsymbol{r}_{\mathrm{neutron}}|\rangle$. For each basis set, the first column refers to the value of the quantity and the second column to the corresponding standard deviation upon averaging. The average neutron N−H bond length is 1.044 Å with a standard deviation of 0.009 Å.

$\langle \boldsymbol{r}_{\mathrm{HAR}}(\mathrm{N-H})\rangle$/Å						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	1.026	0.014	1.030	0.018	1.032	0.016
CCSD	1.022	0.015	1.028	0.017	1.030	0.016
HF	1.031	0.012	1.034	0.012	1.036	0.011
ELMO	1.021	0.010	1.028	0.009	1.031	0.008
BLYP	1.021	0.017	1.017	0.020	1.020	0.019
B3LYP	1.025	0.015	1.023	0.018	1.025	0.017

$\langle \boldsymbol{r}_{\mathrm{HAR}}/\boldsymbol{r}_{\mathrm{neutron}}\rangle$						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.982	0.007	0.986	0.009	0.989	0.008
CCSD	0.979	0.007	0.985	0.008	0.987	0.007
HF	0.987	0.009	0.990	0.008	0.992	0.008
ELMO	0.978	0.004	0.985	0.008	0.988	0.008
BLYP	0.977	0.008	0.974	0.010	0.977	0.009
B3LYP	0.981	0.007	0.980	0.009	0.982	0.008

$\langle\lvert\boldsymbol{r}_{\mathrm{HAR}} - \boldsymbol{r}_{\mathrm{neutron}}\rvert\rangle$/Å						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.018	0.007	0.014	0.009	0.012	0.008
CCSD	0.022	0.007	0.016	0.008	0.014	0.007
HF	0.013	0.009	0.010	0.008	0.008	0.008
ELMO	0.023	0.004	0.016	0.008	0.013	0.008
BLYP	0.023	0.008	0.027	0.010	0.024	0.009
B3LYP	0.019	0.007	0.021	0.009	0.019	0.008

Table A.3: Statistical analysis of the HAR ADPs for non-hydrogen atoms: (i) mean ratios of the diagonal elements $\langle U^{ii}_{\mathrm{HAR}}/U^{ii}_{\mathrm{neutron}}\rangle$, (ii) mean absolute differences of the diagonal terms $\langle|U^{ii}_{\mathrm{HAR}} - U^{ii}_{\mathrm{neutron}}|\rangle$ and (iii) mean absolute differences of the non-diagonal elements $\langle|U^{ij}_{\mathrm{HAR}} - U^{ij}_{\mathrm{neutron}}|\rangle$. For each basis set, the first column refers to the value of the quantity and the second column to the corresponding standard deviation upon averaging.

$\langle U^{ii}_{\mathrm{HAR}}/U^{ii}_{\mathrm{neutron}}\rangle$						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.924	0.081	0.906	0.082	0.907	0.082
CCSD	0.923	0.081	0.906	0.082	0.906	0.082
HF	0.931	0.082	0.917	0.083	0.917	0.083
ELMO	0.924	0.088	0.912	0.086	0.912	0.086
BLYP	0.900	0.081	0.880	0.083	0.881	0.083
B3LYP	0.905	0.081	0.886	0.083	0.886	0.083

$\langle\lvert U^{ii}_{\mathrm{HAR}} - U^{ii}_{\mathrm{neutron}}\rvert\rangle/\mathrm{Å}^2$						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.00053	0.00034	0.00060	0.00037	0.00059	0.00037
CCSD	0.00053	0.00034	0.00060	0.00037	0.00060	0.00037
HF	0.00051	0.00034	0.00056	0.00037	0.00056	0.00037
ELMO	0.00056	0.00039	0.00059	0.00040	0.00059	0.00040
BLYP	0.00062	0.00038	0.00072	0.00040	0.00072	0.00040
B3LYP	0.00060	0.00038	0.00069	0.00040	0.00069	0.00040

$\langle\lvert U^{ij}_{\mathrm{HAR}} - U^{ij}_{\mathrm{neutron}}\rvert\rangle/\mathrm{Å}^2$ with $i \neq j$						
QM method	def2-SVP		def2-TZVP		def2-TZVPP	
MP2	0.000093	0.000077	0.000090	0.000073	0.000090	0.000073
CCSD	0.000090	0.000077	0.000083	0.000076	0.000082	0.000075
HF	0.000104	0.000075	0.000094	0.000076	0.000093	0.000076
ELMO	0.000127	0.000110	0.000109	0.000094	0.000108	0.000093
BLYP	0.000096	0.000078	0.000094	0.000077	0.000094	0.000077
B3LYP	0.000094	0.000078	0.000089	0.000076	0.000088	0.000076

B Appendix to Chapter 5

Table B.1: Inter- and intramolecular contacts in the neutron structure[334] and in the different HAR structures of xylitol (the underling wave functions for all HARs were computed with basis set cc-pVDZ).

	Refinement	D−H	H···A	D···A	D−H···A
O1−H11···O2 (intra)	neutron	0.9960(19)	2.492(2)	2.8654(12)	101.70(12)
	HAR / no embedding	0.961(5)	2.504(6)	2.8686(14)	102.5(4)
	HAR / 4 Å charges	0.968(5)	2.501(6)	2.8687(14)	102.4(4)
	HAR / 8 Å charges	0.968(5)	2.501(6)	2.8687(14)	102.4(4)
	HAR / 4 Å charges & dipoles	0.977(5)	2.499(6)	2.8687(14)	102.2(4)
	HAR / 8 Å charges & dipoles	0.977(5)	2.499(6)	2.8688(14)	102.2(4)
	HAR / 4 Å ELMOs	0.993(5)	2.497(6)	2.8686(14)	101.7(4)
	HAR / 8 Å ELMOs	0.993(5)	2.498(6)	2.8686(14)	101.7(4)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.993(5)	2.498(6)	2.8686(14)	101.7(4)
O1−H11···O5	neutron	0.9960(19)	1.7113(19)	2.6654(12)	159.04(17)
	HAR / no embedding	0.961(5)	1.747(5)	2.6669(13)	159.3(6)
	HAR / 4 Å charges	0.968(5)	1.741(5)	2.6669(13)	159.0(6)
	HAR / 8 Å charges	0.968(5)	1.741(5)	2.6669(13)	159.0(6)
	HAR / 4 Å charges & dipoles	0.977(5)	1.733(5)	2.6669(13)	158.7(5)
	HAR / 8 Å charges & dipoles	0.977(5)	1.733(5)	2.6669(13)	158.7(5)
	HAR / 4 Å ELMOs	0.993(5)	1.718(5)	2.6671(13)	158.7(5)
	HAR / 8 Å ELMOs	0.993(5)	1.717(5)	2.6671(13)	158.7(5)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.993(5)	1.717(5)	2.6671(13)	158.7(5)
O2−H12···O4	neutron	0.978(2)	1.845(2)	2.8218(13)	176.90(17)
	HAR / no embedding	0.955(5)	1.871(5)	2.8242(14)	176.1(5)
	HAR / 4 Å charges	0.960(5)	1.865(5)	2.8241(14)	176.2(5)
	HAR / 8 Å charges	0.960(5)	1.865(5)	2.8241(14)	176.2(5)
	HAR / 4 Å charges & dipoles	0.967(5)	1.858(5)	2.8240(14)	176.4(5)
	HAR / 8 Å charges & dipoles	0.968(5)	1.857(5)	2.8240(14)	176.4(5)
	HAR / 4 Å ELMOs	0.976(5)	1.849(5)	2.8242(14)	176.8(5)
	HAR / 8 Å ELMOs	0.977(5)	1.848(5)	2.8242(14)	176.9(5)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.977(5)	1.849(5)	2.8242(14)	176.8(5)

Continued on next page

Table B.1 – *Continued from previous page*

	Refinement	D−H	H···A	D···A	D−H···A
O3−H13···O1	neutron	0.9863(19)	1.6932(19)	2.6712(13)	170.66(17)
	HAR / no embedding	0.958(5)	1.724(5)	2.6711(13)	169.4(5)
	HAR / 4 Å charges	0.965(5)	1.716(5)	2.6711(13)	169.5(5)
	HAR / 8 Å charges	0.964(5)	1.717(5)	2.6711(13)	169.5(5)
	HAR / 4 Å charges & dipoles	0.974(5)	1.707(5)	2.6711(13)	169.7(5)
	HAR / 8 Å charges & dipoles	0.973(5)	1.708(5)	2.6711(13)	169.7(5)
	HAR / 4 Å ELMOs	0.991(5)	1.691(5)	2.6715(13)	169.7(5)
	HAR / 8 Å ELMOs	0.990(5)	1.692(5)	2.6715(13)	169.7(5)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.990(5)	1.692(5)	2.6715(13)	169.7(5)
O4−H14···O2	neutron	0.970(2)	1.903(2)	2.8475(12)	164.00(17)
	HAR / no embedding	0.957(5)	1.915(5)	2.8478(14)	164.1(4)
	HAR / 4 Å charges	0.961(5)	1.911(5)	2.8477(14)	164.1(4)
	HAR / 8 Å charges	0.961(5)	1.911(5)	2.8477(14)	164.1(4)
	HAR / 4 Å charges & dipoles	0.968(5)	1.904(5)	2.8478(14)	164.2(4)
	HAR / 8 Å charges & dipoles	0.968(5)	1.904(5)	2.8477(14)	164.2(4)
	HAR / 4 Å ELMOs	0.975(5)	1.898(5)	2.8478(14)	164.1(4)
	HAR / 8 Å ELMOs	0.975(5)	1.898(5)	2.8478(14)	164.1(4)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.975(5)	1.898(5)	2.8478(14)	164.1(4)
O5−H15···O3	neutron	0.986(2)	1.716(2)	2.7000(13)	176.22(17)
	HAR / no embedding	0.947(6)	1.754(6)	2.7001(13)	178.3(5)
	HAR / 4 Å charges	0.953(5)	1.748(5)	2.7002(13)	178.0(5)
	HAR / 8 Å charges	0.953(5)	1.748(5)	2.7002(13)	178.0(5)
	HAR / 4 Å charges & dipoles	0.964(5)	1.737(5)	2.7001(13)	177.5(5)
	HAR / 8 Å charges & dipoles	0.964(5)	1.736(5)	2.7001(13)	177.5(5)
	HAR / 4 Å ELMOs	0.979(5)	1.723(5)	2.7002(13)	176.9(5)
	HAR / 8 Å ELMOs	0.980(5)	1.722(5)	2.7003(13)	176.8(5)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.979(5)	1.722(5)	2.7003(13)	176.8(5)
C1−H1B···O1	neutron	1.098(2)	2.5088(19)	3.3229(11)	130.00(18)
	HAR / no embedding	1.071(5)	2.528(5)	3.3206(16)	130.1(4)
	HAR / 4 Å charges	1.074(5)	2.525(5)	3.3206(16)	130.1(3)
	HAR / 8 Å charges	1.074(5)	2.525(5)	3.3206(16)	130.1(3)
	HAR / 4 Å charges & dipoles	1.078(5)	2.520(5)	3.3207(16)	130.3(3)
	HAR / 8 Å charges & dipoles	1.079(5)	2.520(5)	3.3207(16)	130.2(3)
	HAR / 4 Å ELMOs	1.081(5)	2.518(5)	3.3207(16)	130.3(4)
	HAR / 8 Å ELMOs	1.082(5)	2.517(5)	3.3207(16)	130.2(4)
	4 Å ELMOs + MM (4 Å - 8 Å)	1.082(5)	2.517(5)	3.3207(16)	130.2(4)
C4−H4···O1	neutron	1.1028(18)	2.3261(18)	3.2928(11)	145.25(14)
	HAR / no embedding	1.078(4)	2.351(5)	3.2959(16)	145.5(3)
	HAR / 4 Å charges	1.083(4)	2.345(5)	3.2958(16)	145.5(3)
	HAR / 8 Å charges	1.083(4)	2.345(5)	3.2958(16)	145.6(3)
	HAR / 4 Å charges & dipoles	1.091(4)	2.338(5)	3.2958(16)	145.6(3)
	HAR / 8 Å charges & dipoles	1.091(4)	2.338(5)	3.2958(16)	145.6(3)
	HAR / 4 Å ELMOs	1.094(4)	2.334(5)	3.2956(16)	145.7(3)
	HAR / 8 Å ELMOs	1.095(4)	2.333(5)	3.2956(16)	145.7(3)
	4 Å ELMOs + MM (4 Å - 8 Å)	1.095(4)	2.333(5)	3.2956(16)	145.7(3)

Table B.2: Inter- and intramolecular contacts in the neutron structure[334] and in the different HAR structures of xylitol (the underling wave functions for all HARs were computed with basis set cc-pVTZ).

	Refinement	D−H	H···A	D···A	D−H···A
O1−H11···O2 (intra)	neutron	0.9960(19)	2.492(2)	2.8654(12)	101.70(12)
	no embedding	0.963(5)	2.506(6)	2.8690(14)	102.2(4)
	4 Å charges	0.969(5)	2.503(6)	2.8691(14)	102.2(4)
	8 Å charges	0.968(5)	2.504(6)	2.8691(14)	102.2(4)
	4 Å charges & dipoles	0.976(5)	2.500(5)	2.8692(14)	102.2(3)
	8 Å charges & dipoles	0.976(5)	2.500(5)	2.8693(14)	102.2(3)
	4 Å ELMOs	0.994(5)	2.501(5)	2.8690(14)	101.5(3)
	8 Å ELMOs	0.994(5)	2.501(5)	2.8690(14)	101.5(3)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.994(5)	2.501(5)	2.8690(14)	101.5(3)
O1−H11···O5	neutron	0.9960(19)	1.7113(19)	2.6654(12)	159.04(17)
	no embedding	0.963(5)	1.741(5)	2.6667(13)	160.0(5)
	4 Å charges	0.969(5)	1.738(5)	2.6667(13)	159.6(5)
	8 Å charges	0.968(5)	1.737(5)	2.6666(13)	159.6(5)
	4 Å charges & dipoles	0.976(5)	1.732(5)	2.6665(13)	159.1(5)
	8 Å charges & dipoles	0.976(5)	1.731(5)	2.6666(13)	159.1(5)
	4 Å ELMOs	0.994(5)	1.714(5)	2.6669(13)	159.3(5)
	8 Å ELMOs	0.994(5)	1.714(5)	2.6669(13)	159.3(5)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.994(5)	1.714(5)	2.6669(13)	159.4(5)
O2−H12···O4	neutron	0.978(2)	1.845(2)	2.8218(13)	176.90(17)
	no embedding	0.950(5)	1.875(5)	2.8241(14)	176.8(5)
	4 Å charges	0.954(5)	1.871(5)	2.8241(14)	176.9(5)
	8 Å charges	0.954(5)	1.871(5)	2.8241(14)	176.9(5)
	4 Å charges & dipoles	0.961(5)	1.864(5)	2.8240(14)	177.0(5)
	8 Å charges & dipoles	0.962(5)	1.863(5)	2.8240(14)	177.1(5)
	4 Å ELMOs	0.971(5)	1.854(5)	2.8241(14)	177.5(5)
	8 Å ELMOs	0.972(5)	1.853(5)	2.8241(14)	177.6(5)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.972(5)	1.853(5)	2.8241(14)	177.6(5)
O3−H13···O1	neutron	0.9863(19)	1.6932(19)	2.6712(13)	170.66(17)
	no embedding	0.968(5)	1.713(5)	2.6710(13)	169.8(5)
	4 Å charges	0.973(5)	1.708(5)	2.6710(13)	170.0(5)
	8 Å charges	0.973(5)	1.708(5)	2.6709(13)	169.9(5)
	4 Å charges & dipoles	0.980(5)	1.700(5)	2.6709(13)	170.2(5)
	8 Å charges & dipoles	0.980(5)	1.700(5)	2.6709(13)	170.1(5)
	4 Å ELMOs	1.001(5)	1.680(5)	2.6713(13)	170.2(5)
	8 Å ELMOs	1.001(5)	1.680(5)	2.6713(13)	170.1(5)
	4 Å ELMOs + MM (4 Å - 8 Å)	1.001(5)	1.680(5)	2.6712(13)	170.2(5)
O4−H14···O2	neutron	0.970(2)	1.903(2)	2.8475(12)	164.00(17)
	no embedding	0.954(5)	1.917(5)	2.8476(14)	164.5(4)
	4 Å charges	0.957(5)	1.914(5)	2.8476(14)	164.5(4)
	8 Å charges	0.957(5)	1.914(5)	2.8476(14)	164.5(4)
	4 Å charges & dipoles	0.963(5)	1.907(5)	2.8476(14)	164.8(4)
	8 Å charges & dipoles	0.963(5)	1.907(5)	2.8476(14)	164.8(4)
	4 Å ELMOs	0.971(5)	1.900(5)	2.8476(14)	164.6(4)
	8 Å ELMOs	0.970(5)	1.900(5)	2.8476(14)	164.7(4)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.970(5)	1.900(5)	2.8476(14)	164.7(4)

Continued on next page

Table B.2 – *Continued from previous page*

	Refinement	D−H	H···A	D···A	D−H···A
O5−H15···O3	neutron	0.986(2)	1.716(2)	2.7000(13)	176.22(17)
	no embedding	0.941(5)	1.759(5)	2.7004(13)	178.3(6)
	4 Å charges	0.947(5)	1.754(5)	2.7003(13)	177.9(6)
	8 Å charges	0.947(5)	1.754(5)	2.7003(13)	177.9(6)
	4 Å charges & dipoles	0.957(5)	1.744(5)	2.7002(13)	177.3(5)
	8 Å charges & dipoles	0.958(5)	1.743(5)	2.7002(13)	177.3(5)
	4 Å ELMOs	0.975(5)	1.726(5)	2.7003(13)	176.7(5)
	8 Å ELMOs	0.976(5)	1.725(5)	2.7004(13)	176.6(5)
	4 Å ELMOs + MM (4 Å - 8 Å)	0.976(5)	1.725(5)	2.7003(13)	176.7(5)
C1−H1B···O1	neutron	1.098(2)	2.5088(19)	3.3229(11)	130.00(18)
	no embedding	1.077(4)	2.530(4)	3.3206(16)	129.5(3)
	4 Å charges	1.081(4)	2.526(4)	3.3206(16)	129.6(3)
	8 Å charges	1.081(4)	2.526(4)	3.3206(16)	129.6(3)
	4 Å charges & dipoles	1.085(4)	2.520(4)	3.3206(16)	129.7(3)
	8 Å charges & dipoles	1.086(4)	2.520(4)	3.3206(16)	129.7(3)
	4 Å ELMOs	1.087(4)	2.518(4)	3.3207(16)	129.8(3)
	8 Å ELMOs	1.089(4)	2.517(4)	3.3207(16)	129.8(3)
	4 Å ELMOs + MM (4 Å - 8 Å)	1.089(4)	2.517(4)	3.3207(16)	129.8(3)
C4−H4···O1	neutron	1.1028(18)	2.3261(18)	3.2928(11)	145.25(14)
	no embedding	1.082(4)	2.346(5)	3.2956(16)	145.6(3)
	4 Å charges	1.087(4)	2.341(5)	3.2955(16)	145.6(3)
	8 Å charges	1.087(4)	2.341(5)	3.2955(16)	145.6(3)
	4 Å charges & dipoles	1.095(4)	2.332(4)	3.2954(16)	145.7(3)
	8 Å charges & dipoles	1.097(4)	2.331(4)	3.2954(16)	145.7(3)
	4 Å ELMOs	1.098(4)	2.329(5)	3.2954(16)	145.8(3)
	8 Å ELMOs	1.099(4)	2.329(5)	3.2954(16)	145.7(3)
	4 Å ELMOs + MM (4 Å - 8 Å)	1.099(4)	2.329(5)	3.2953(16)	145.7(3)

C Appendix to Chapter 9

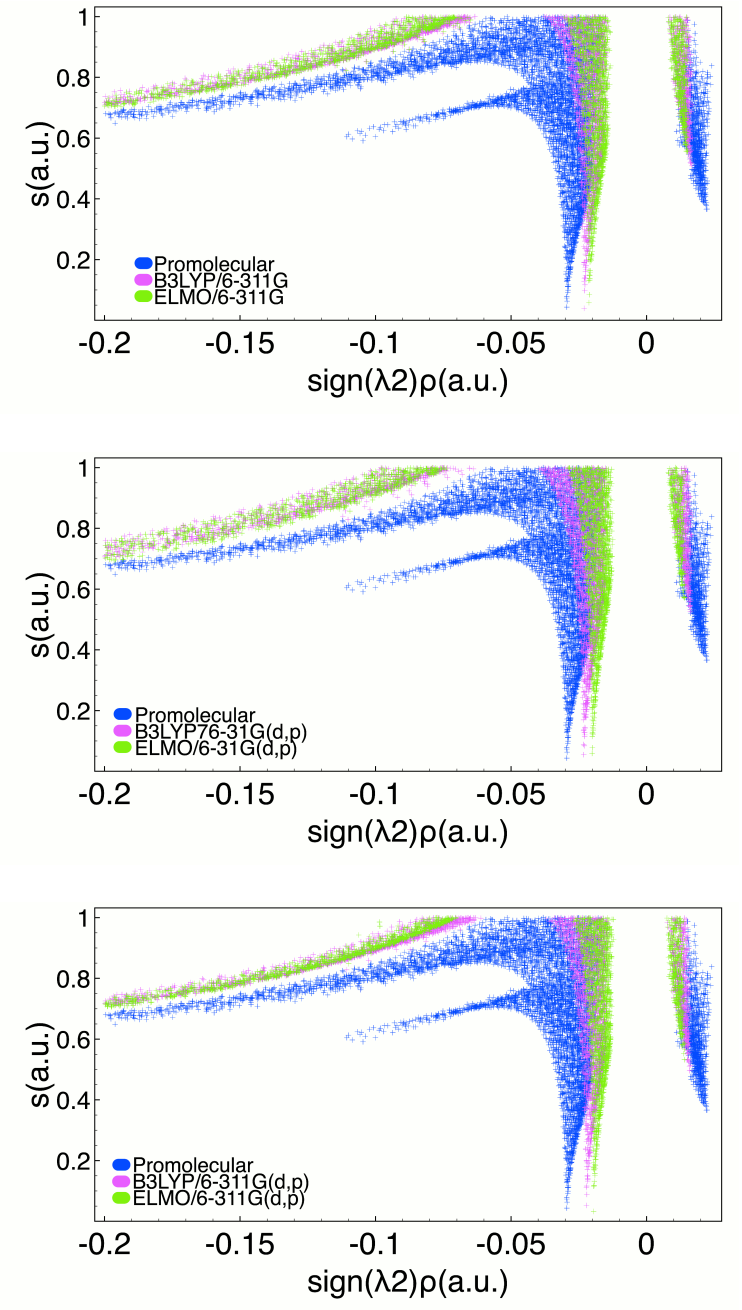


Figure C.1: Strong hydrogen bond between Tyr2 and Phe5 in Leu-enkephalin:[314] 2D RDG plots obtained at promolecular-NCI, NCI-B3LYP and NCI-ELMO levels for the basis sets 6-311G, 6-31G(d,p) and 6-311G(d,p) as indicated in the legends.

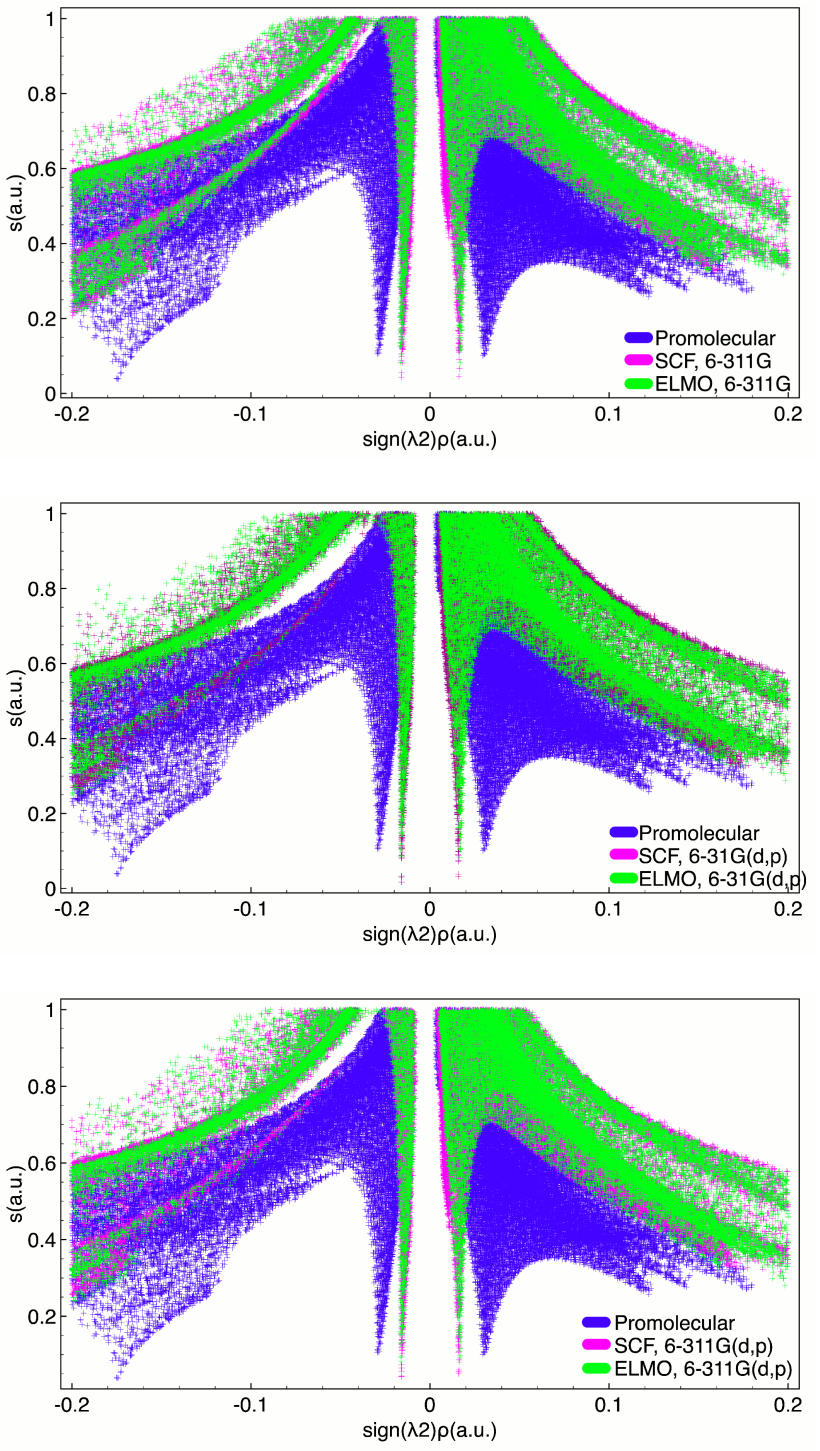


Figure C.2: Weak hydrogen bond between Phe2 and Gly3 in lactoferrampin (PDB code: 2MD3): 2D RDG plots obtained at promolecular-NCI, NCI-B3LYP and NCI-ELMO levels for the basis sets 6-311G, 6-31G(d,p) and 6-311G(d,p) as indicated in the legends.

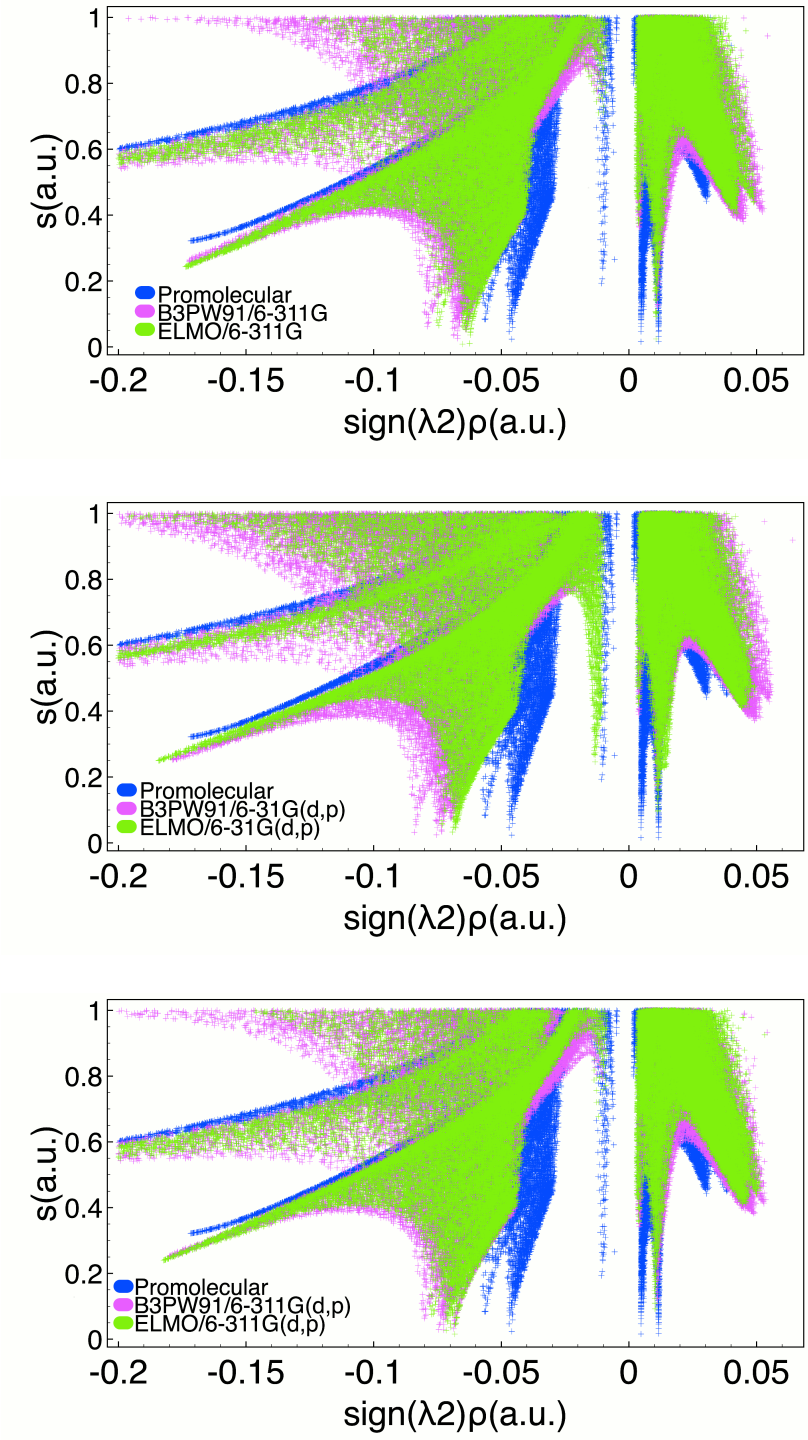


Figure C.3: Interactions between the Zn^{2+} ion and the coordinating residues (Cys3, Cys6, Cys16 and His11) in an HIV Zinc fingerlike domain (PDB code: 2ZNF): 2D RDG plots obtained at promolecular-NCI, NCI-B3PW91 and NCI-ELMO levels for the basis sets 6-311G, 6-31G(d,p) and 6-311G(d,p) as indicated in the legends.

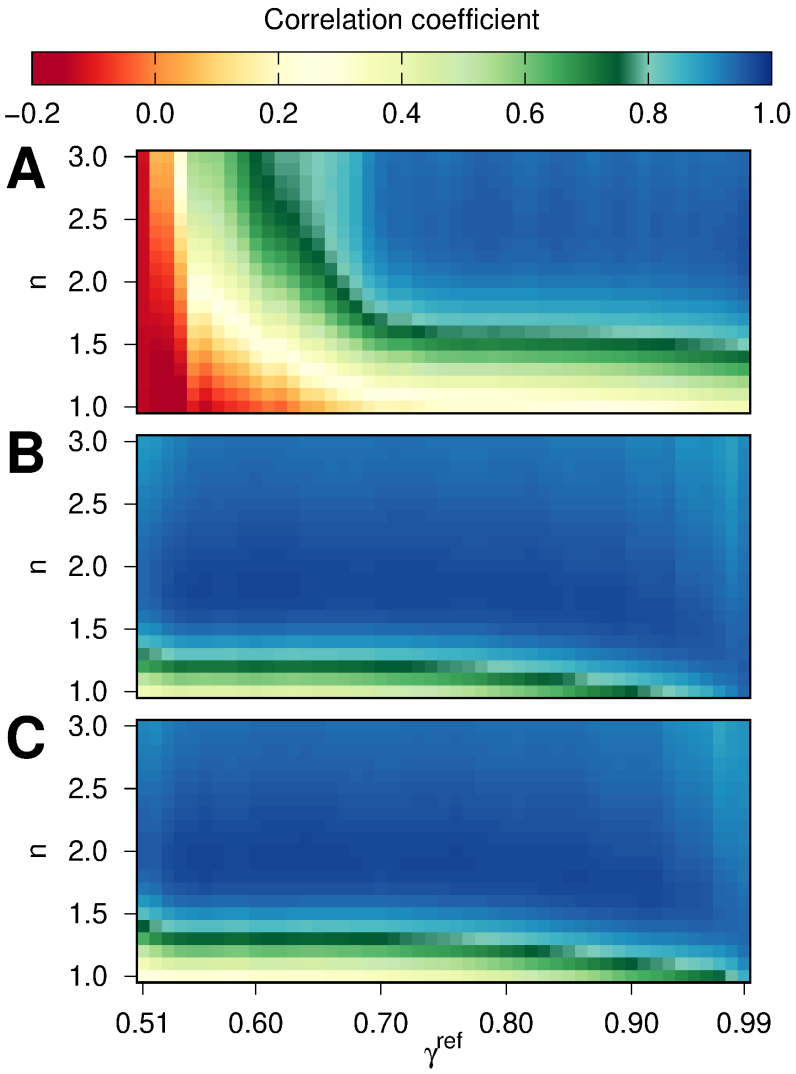


Figure C.4: Heatmaps showing the variation of the correlation coefficients according to different values of the integral exponents n and electron density thresholds γ^{ref}. Each correlation coefficient gives the correlation between the interaction energies and the NCI integral values for the S66 dataset. The following densities were used to compute the NCI integrals are based on the following electron density approximations: ((A) promolecular, (B) ELMO/6-31G (monomer approximation) and (C) ELMO/6-31G (dimer approximation) electron densities.

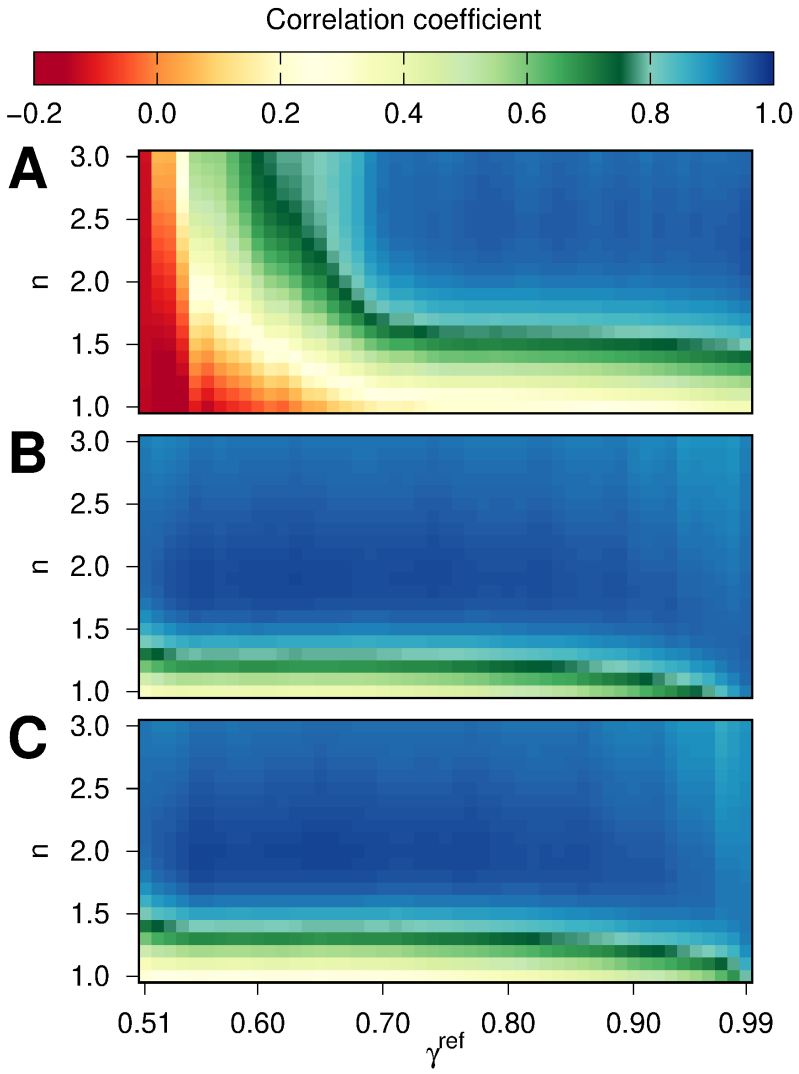


Figure C.5: Heatmaps showing the variation of the correlation coefficients according to different values of the integral exponents n and electron density thresholds γ^{ref}. Each correlation coefficient gives the correlation between the interaction energies and the NCI integral values for the S66 dataset. The following densities were used to compute the NCI integrals are based on the following electron density approximations: ((A) promolecular, (B) ELMO/6-311G (monomer approximation) and (C) ELMO/6-311G (dimer approximation) electron densities.

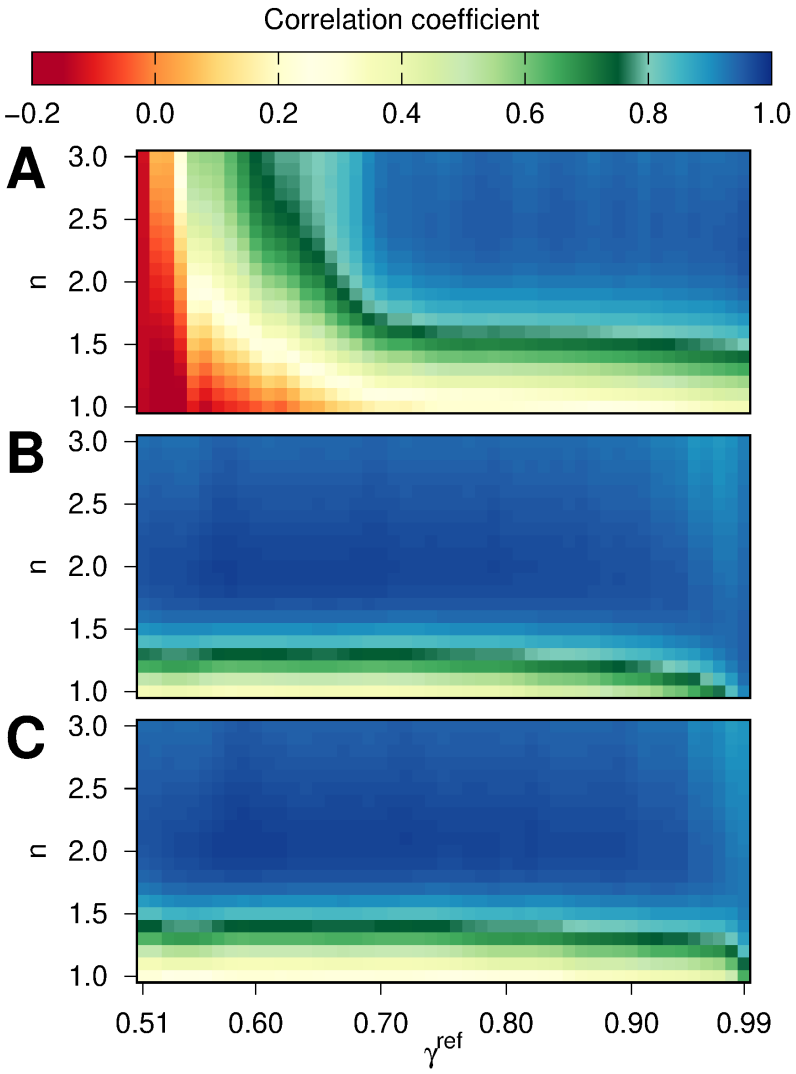


Figure C.6: Heatmaps showing the variation of the correlation coefficients according to different values of the integral exponents n and electron density thresholds γ^{ref}. Each correlation coefficient gives the correlation between the interaction energies and the NCI integral values for the S66 dataset. The following densities were used to compute the NCI integrals are based on the following electron density approximations: ((A) promolecular, (B) ELMO/6-311G(d,p) (monomer approximation) and (C) ELMO/6-311G(d,p) (dimer approximation) electron densities.

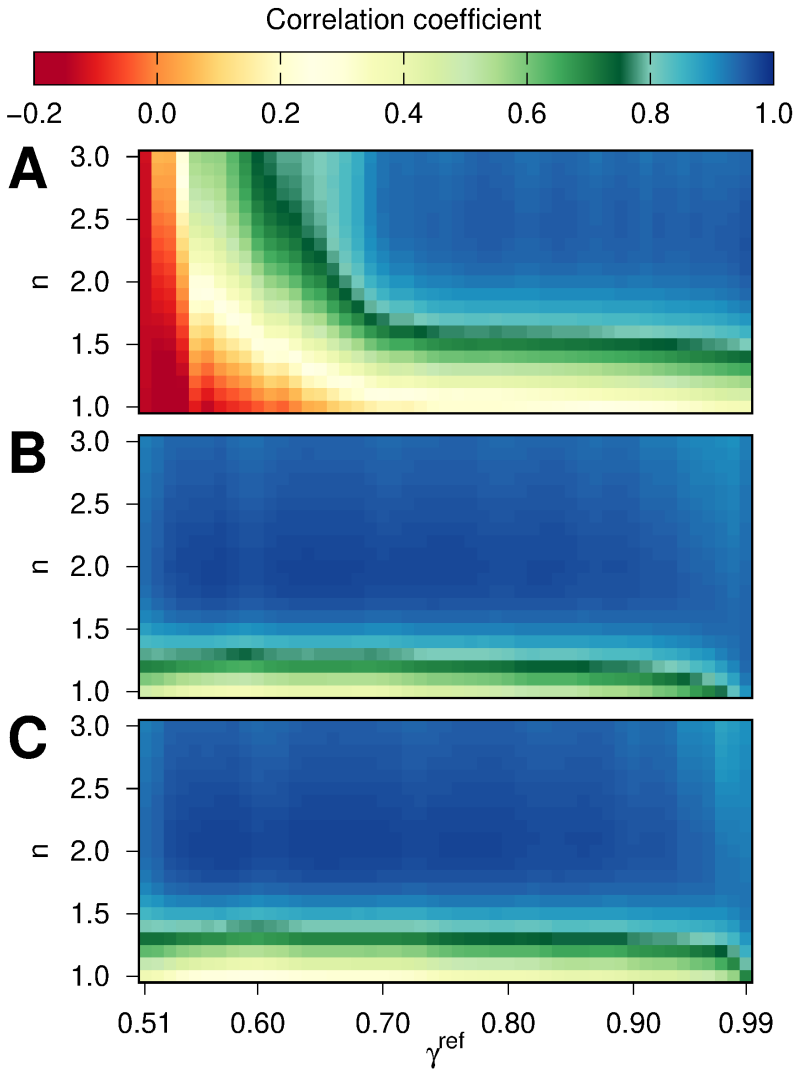


Figure C.7: Heatmaps showing the variation of the correlation coefficients according to different values of the integral exponents n and electron density thresholds γ^{ref}. Each correlation coefficient gives the correlation between the interaction energies and the NCI integral values for the S66 dataset. The following densities were used to compute the NCI integrals are based on the following electron density approximations: ((A) promolecular, (B) ELMO/cc-pVDZ (monomer approximation) and (C) ELMO/cc-pVDZ (dimer approximation) electron densities.

D Appendix to Chapter 10

D.1 Model molecules for the computation of the ELMOs for the halogenated tyrosine residues

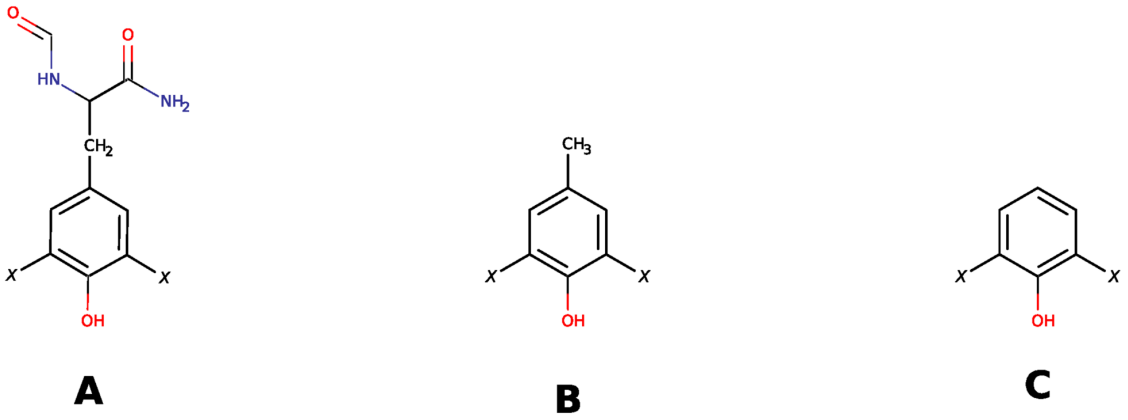


Figure D.1: Model molecules used for the computation of tailor-made ELMOs for the halogenated tyrosine residues (X = Cl, Br).

D.2 Additional Figures for the analysis of the non-covalent interactions in Leu-enkephalin and in the synthetic peptide 1DEP

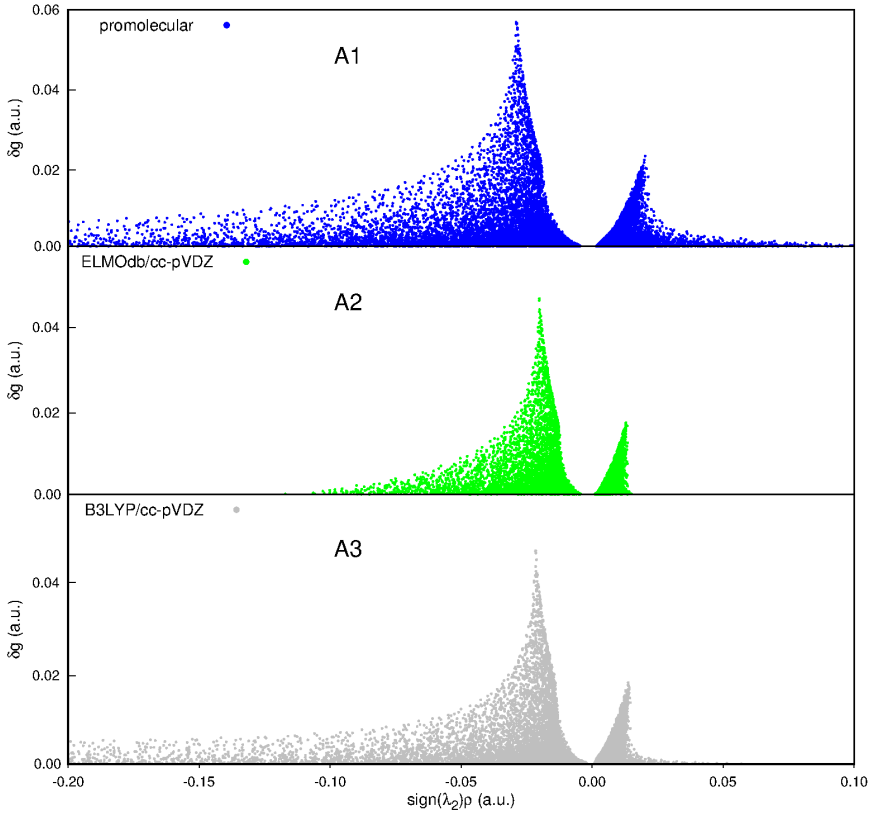


Figure D.2: Comparison between the 2D IGM plots obtained at (A1) promolecular, (A2) ELMOdb/cc-pVDZ and (A3) B3LYP/cc-pVDZ levels for the local hydrogen-bond between residues Tyr1 and Phe4 in Leu-enkephalin.

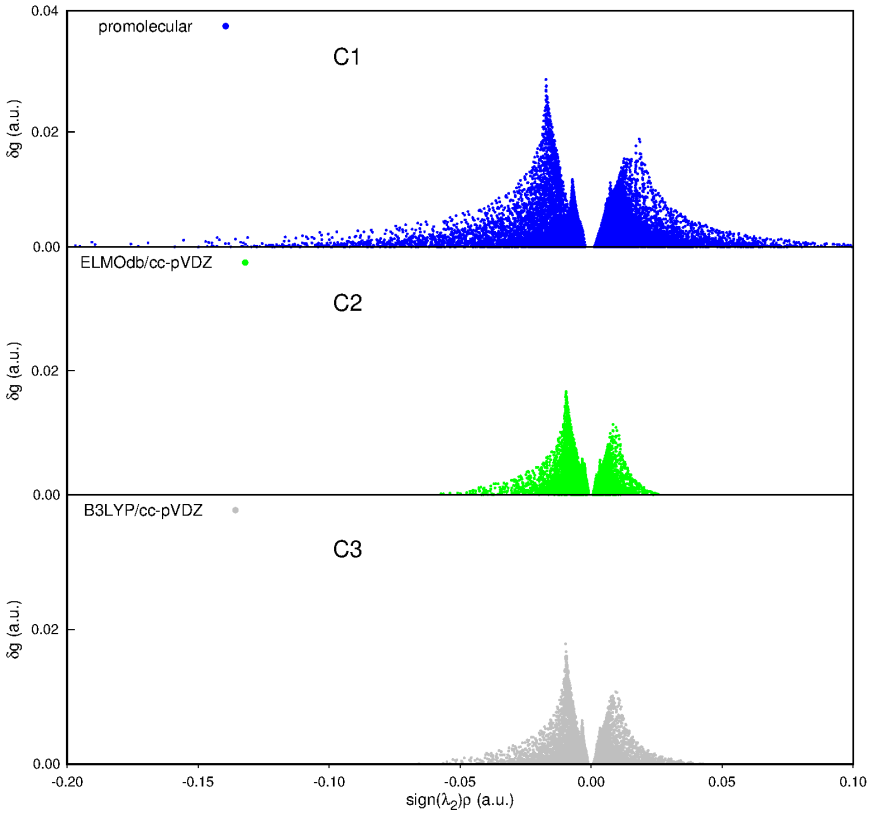


Figure D.3: Comparison between the 2D IGM plots obtained at (C1) promolecular, (C2) ELMOdb/cc-pVDZ and (C3) B3LYP/cc-pVDZ levels for the T-shaped $\pi-\pi$ stacking between Tyr1 and Phe4 in Leu-enkephalin.

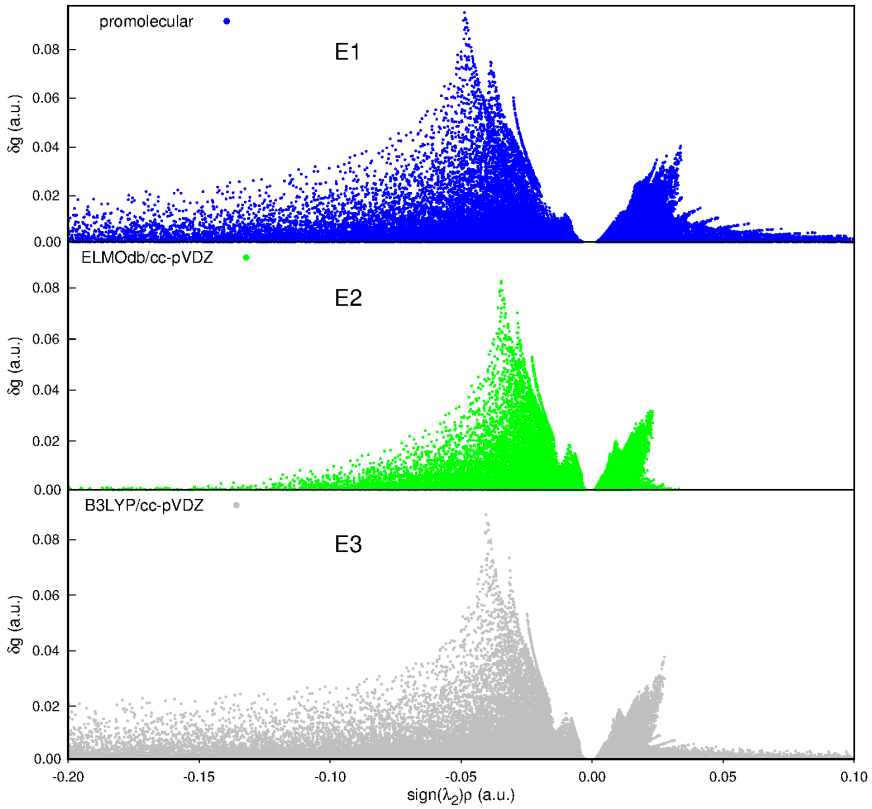


Figure D.4: Comparison between the 2D IGM plots obtained at (E1) promolecular, (E2) ELMOdb/cc-pVDZ and (E3) B3LYP/cc-pVDZ levels for the multiple hydrogen bond interaction between charged residues Asp4, Arg1 and Arg11 in polypeptide 1DEP.

D.3 Additional Figures for the analysis of the non-covalent interactions in halogenated peptide dimers

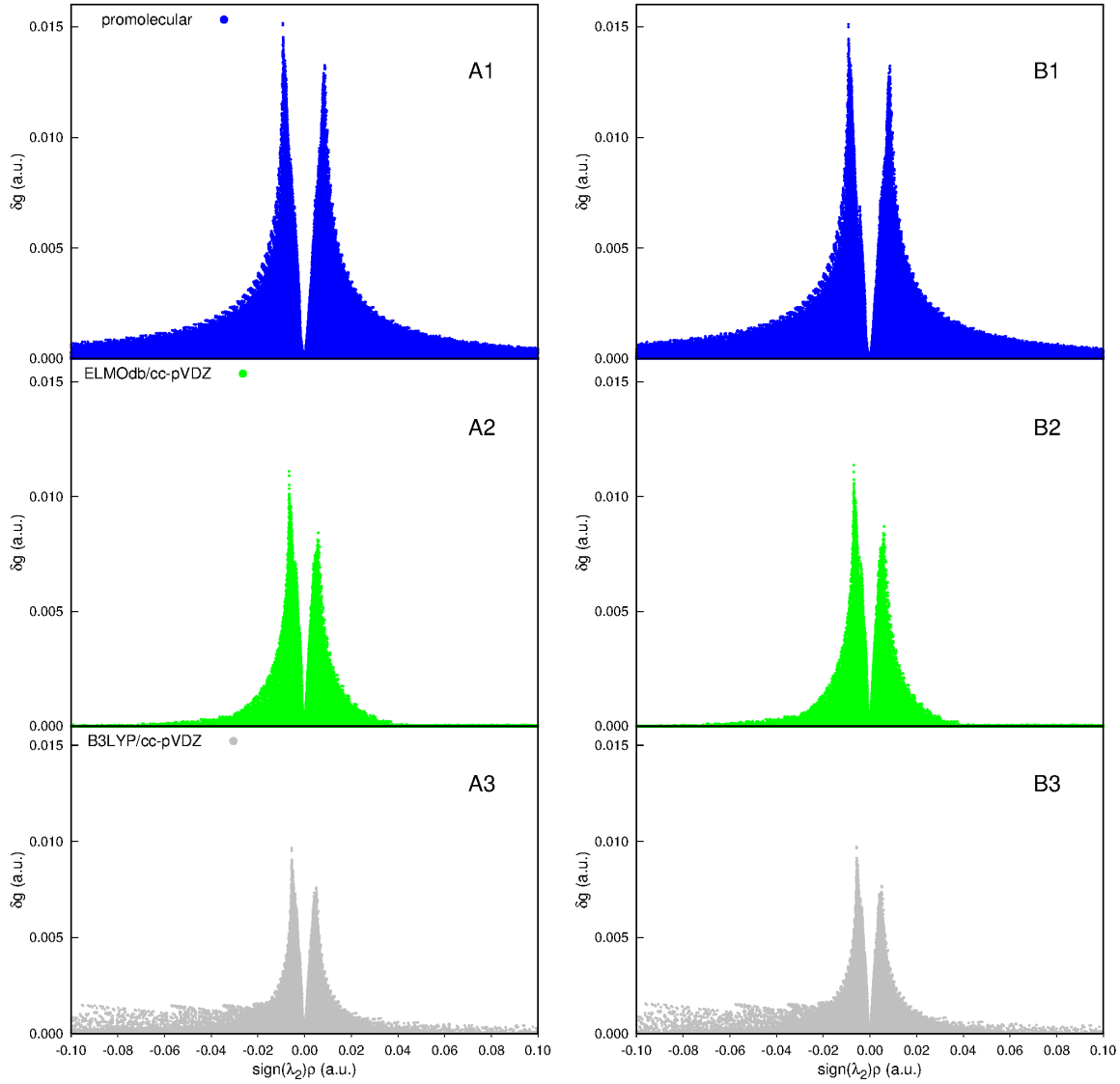


Figure D.5: Comparison between the 2D IGM plots obtained at (1) promolecular (2) ELMOdb/cc-pVDZ and (3) B3LYP/cc-pVDZ levels for the $\pi - \pi$ interactions in the dimers of the (A) brominated and (B) debrominated polypeptides.

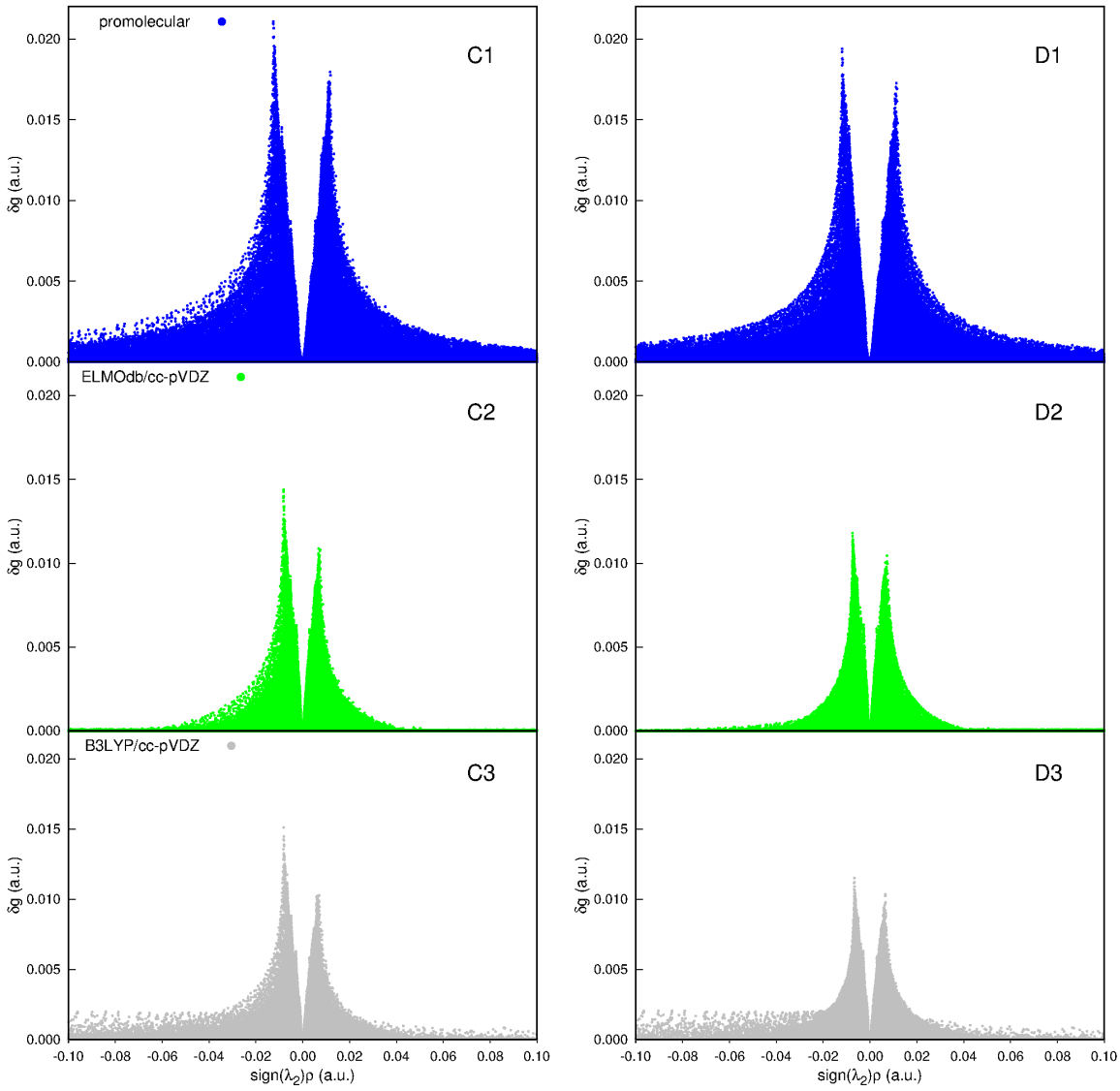


Figure D.6: Comparison between the 2D IGM plots obtained at (1) promolecular (2) ELMOdb/cc-pVDZ and (3) B3LYP/cc-pVDZ, levels for the $\pi - \pi$ interactions in the dimers of the (C) chlorinated and (D) dechlorinated polypeptides.

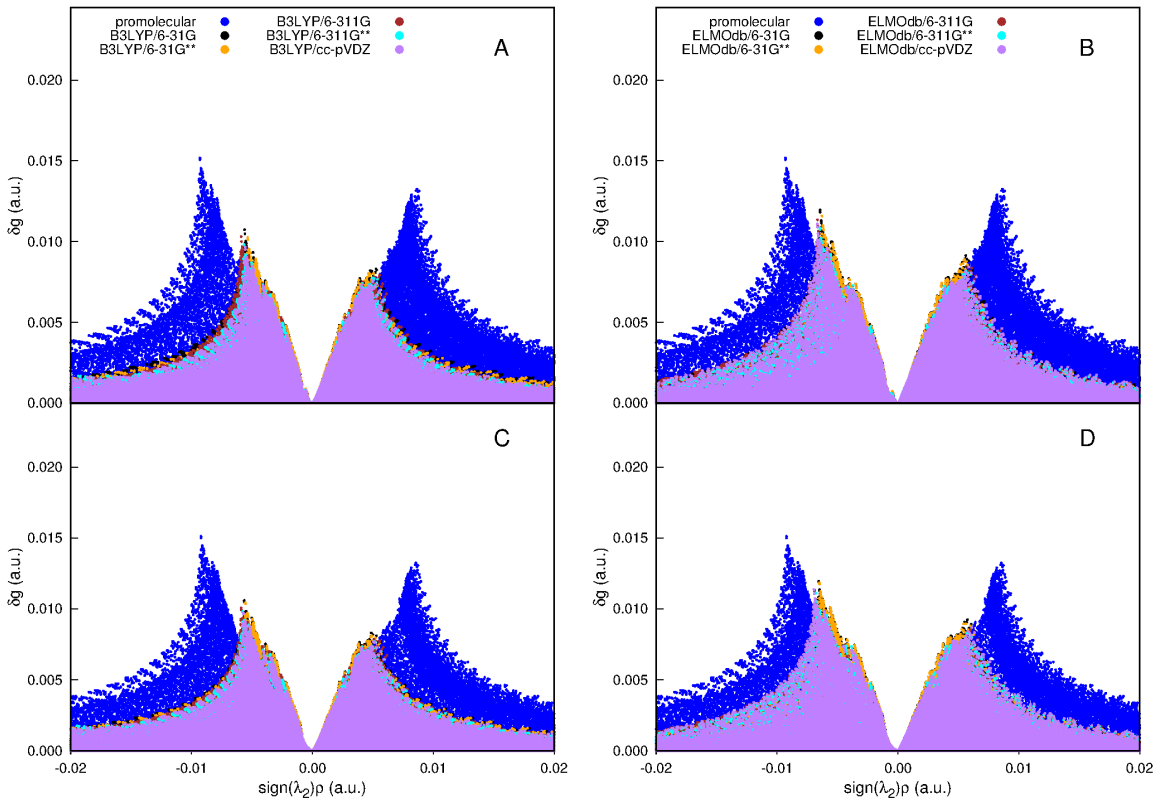


Figure D.7: Basis set dependence of the 2D IGM plots associated with the $\pi - \pi$ interactions in the dimers of the brominated (A and B) and debrominated (C and D) polypeptides. (A) and (C) refer to promolecular and DFT calculations, while (B) and (D) refer to promolecular and ELMOdb computations.

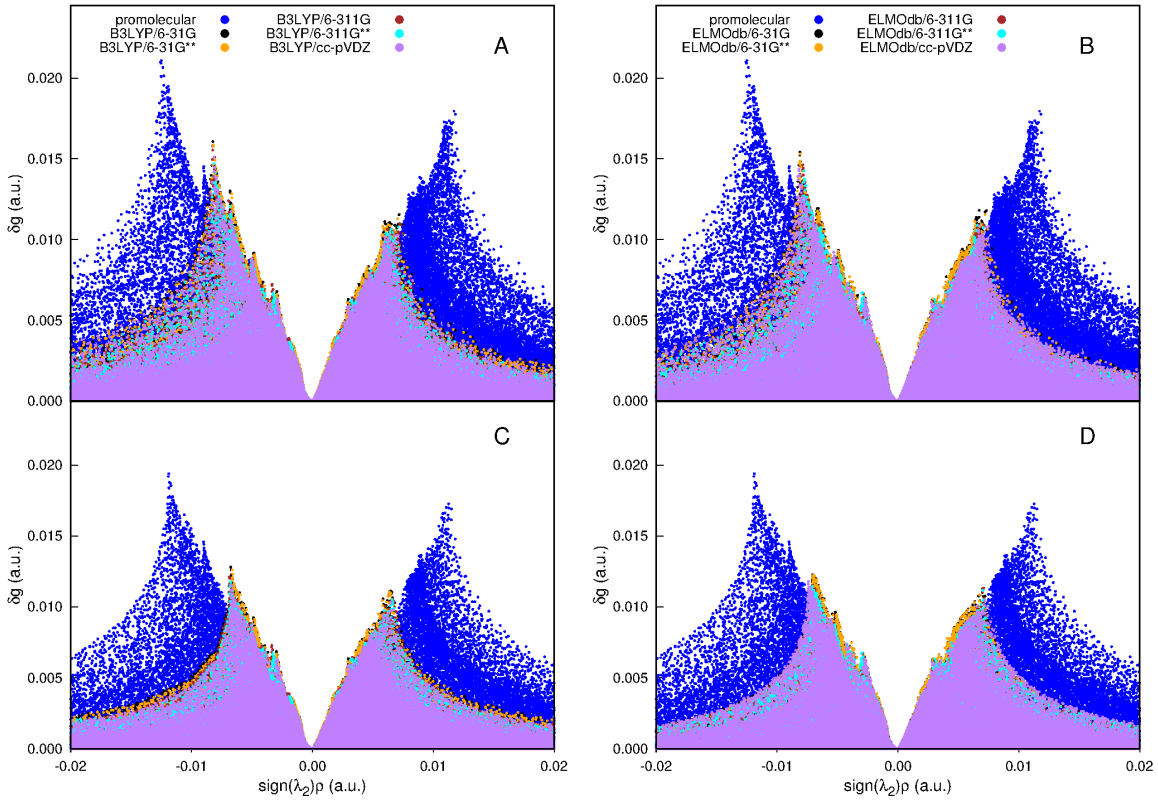


Figure D.8: Basis set dependence of the 2D IGM plots associated with the $\pi - \pi$ interactions in the dimers of the chlorinated (A and B) and dechlorinated (C and D) polypeptides. (A) and (C) refer to promolecular and DFT calculations, while (B) and (D) refer to promolecular and ELMOdb computations.

List of Publications

Publications that are included in this thesis:

1. D. Arias-Olivares, E. K. Wieduwilt, J. Contreras-García, A. Genoni, "NCI-ELMO: A new method to quickly and accurately detect noncovalent interactions in biosystems", *J. Chem. Theory Comput.* **2019**, *15*, 6456–6470

2. L. A. Malaspina, E. K. Wieduwilt, J. Bergmann, F. Kleemiss, B. Meyer, M. F. Ruiz-López, R. Pal, E. Hupf, J. Beckmann, R. O. Piltz, A. J. Edwards, S. Grabowsky, A Genoni, "Fast and accurate quantum crystallography: from small to large, from light to heavy", *J. Phys. Chem. Lett.* **2019**, *10*, 6973–6982

3. E. K. Wieduwilt, G. Macetti, L. A. Malaspina, D. Jayatilaka, S. Grabowsky, A. Genoni, "Post-Hartree-Fock methods for Hirshfeld atom refinement: are they necessary? Investigation of a strongly hydrogen-bonded molecular crystal", *J. Mol. Struct.* **2020**, *1209*, 127934

4. E. K. Wieduwilt, G. Macetti, A. Genoni, "Climbing Jacob's ladder of structural refinement: Introduction of a localized molecular orbital-based embedding for accurate X-ray determinations of hydrogen atom positions", *J. Phys. Chem. Lett.* **2021**, *12*, 463–471

5. E. K. Wieduwilt, J.-C. Boisson, G. Terraneo, E. Hénon, A. Genoni, "A Step toward the quantification of noncovalent interactions in large biological systems: The independent gradient model-extremely localized molecular orbital approach", *J. Chem. Inf. Model.* **2021**, *61*, 795–809

Other publications that were published in the framework of this thesis:

6. G. Macetti, E. K. Wieduwilt, X. Assfeld, A. Genoni, "Localized molecular orbital-based embedding scheme for correlated methods", *J. Chem. Theory Comput.* **2020**, *16*, 3578–3596

7. E. K. Wieduwilt, G. Macetti, R. Scatena, P. Macchi, A. Genoni, "Extending libraries of extremely localized molecular orbitals to metal organic frameworks: A preliminary investigation", *Crystals* **2021**, *11*, 207

8. G. Macetti, E. K. Wieduwilt, A. Genoni, "QM/ELMO: A multi-purpose fully quantum mechanical embedding scheme based on extremely localized molecular orbitals", *J. Phys. Chem. A* **2021**, *125*, 2709–2726

Publication related to previous work:

9. F. Kleemiss, E. K. Wieduwilt, E. Hupf, M. W. Shi, S. G. Stewart, D. Jayatilaka, M. J. Turner, K. Sugimoto, E. Nishibori, T. Schirmeister, et al., "Similarities and differences between crystal and enzyme environmental effects on the electron density of drug molecules", *Chem. Eur. J.* **2021**, *27*, 3407

List of Figures

List of Tables

Bibliography

[1] P. A. M. Dirac, R. H. Fowler, "Quantum mechanics of many-electron systems", *Proc. R. Soc. A* **1929**, *123*, 714–733.

[2] J. A. Pople, "Nobel Lecture: Quantum chemical models", *Rev. Mod. Phys.* **1999**, *71*, 1267–1274.

[3] W. Kohn, "Nobel Lecture: Electronic structure of matter—wave functions and density functionals", *Rev. Mod. Phys.* **1999**, *71*, 1253–1266.

[4] E. Schrödinger, "Quantisierung als Eigenwertproblem", *Ann. Phys. (Berl.)* **1926**, *384*, 361–376.

[5] M. Born, R. Oppenheimer, "Zur Quantentheorie der Molekeln", *Ann. Phys. (Berl.)* **1927**, *389*, 457–484.

[6] A. Szabo, N. S. Ostlund, *Modern quantum chemistry: introduction to advanced electronic structure theory*, Dover Publications Inc., **1996**, pp. 39–152.

[7] I. N. Levine, *Quantum chemistry*, *Vol. 7*, Pearson Prentice Hall Upper Saddle River, NJ, **2014**, pp. 180–181.

[8] S. Grabowsky, A. Genoni, S. P. Thomas, D. Jayatilaka, "The advent of quantum crystallography: form and structure factors from quantum mechanics for advanced structure refinement and wavefunction fitting", **2020**, (Eds.: D. Mingos, P. R. Raithby), 65–144.

[9] J. C. Slater, "The theory of complex spectra", *Phys. Rev.* **1929**, *34*, 1293.

[10] R. S. Mulliken, "Electronic structures of polyatomic molecules and valence. II. General considerations", *Phys. Rev.* **1932**, *41*, 49.

[11] D. R. Hartree in Math. Proc. Camb. Philos. Soc. *Vol. 24*, Cambridge University Press, **1928**, pp. 89–110.

[12] D. R. Hartree in Math. Proc. Camb. Philos. Soc. *Vol. 24*, Cambridge University Press, **1928**, pp. 111–132.

[13] V. Fock, "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems", *Z. Phys* **1930**, *61*, 126–148.

[14] C. C. J. Roothaan, "New developments in molecular orbital theory", *Rev. Mod. Phys* **1951**, *23*, 69.

[15] G. Hall, "The molecular orbital theory of chemical valency VIII. A method of calculating ionization potentials", *Proc. R. Soc. A* **1951**, *205*, 541–552.

[16] D. Feller, "The role of databases in support of computational chemistry calculations", *J. Comput. Chem.* **1996**, *17*, 1571–1586.

[17] K. L. Schuchardt, B. T. Didier, T. Elsethagen, et al., "Basis set exchange: a community database for computational sciences", *J. Chem. Inf. Model.* **2007**, *47*, 1045–1052.

[18] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson, T. L. Windus, "New basis set exchange: An open, up-to-date resource for the molecular sciences community", *J. Chem. Inf. Model.* **2019**, *59*, 4814–4820.

[19] R. A. Friesner, "*Ab initio* quantum chemistry: Methodology and applications", *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6648–6653.

[20] C. Møller, M. S. Plesset, "Note on an approximation treatment for many-electron systems", *Phys. Rev.* **1934**, *46*, 618.

[21] M. Head-Gordon, J. A. Pople, M. J. Frisch, "MP2 energy evaluation by direct methods", *Chem. Phys. Lett.* **1988**, *153*, 503–506.

[22] F. Coester, "Bound states of a many-particle system", *Nucl. Phys.* **1958**, *7*, 421–424.

[23] F. Coester, H. Kümmel, "Short-range correlations in nuclear wave functions", *Nucl. Phys.* **1960**, *17*, 477–485.

[24] J. Čížek, "On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods", *J. Chem. Phys.* **1966**, *45*, 4256–4266.

[25] R. J. Bartlett, "Coupled-cluster approach to molecular structure and spectra: a step toward predictive quantum chemistry", *J. Phys. Chem.* **1989**, *93*, 1697–1708.

[26] D. J. Tozer in *European Summerschool in Quantum Chemistry 2019 Book II*, (Eds.: S. Reine, T. Saue), The Exchange-Correlation Energy, **2019**.

[27] J. Toulouse in *European Summerschool in Quantum Chemistry 2019*, (Eds.: S. Reine, T. Saue), Basic density-functional theory, **2019**.

[28] P. Hohenberg, W. Kohn, "Inhomogeneous electron gas", *Phys. Rev.* **1964**, *136*, B864.

[29] W. Kohn, L. J. Sham, "Self-consistent equations including exchange and correlation effects", *Phys. Rev.* **1965**, *140*, A1133.

[30] S. Kurth, J. P. Perdew, "Role of the exchange–correlation energy: Nature's glue", *Int. J. Quantum Chem.* **2000**, *77*, 814–818.

[31] J. P. Perdew, K. Schmidt in AIP Conf. Proc. *Vol. 577*, American Institute of Physics, **2001**, pp. 1–20.

[32] J. P. Perdew, A. Ruzsinszky, L. A. Constantin, J. Sun, G. I. Csonka, "Some fundamental issues in ground-state density functional theory: A guide for the perplexed", *J. Chem. Theory Comput.* **2009**, *5*, 902–908.

[33] J. P. Perdew, K. Burke, M. Ernzerhof, "Generalized gradient approximation made simple", *Phys. Rev. Lett.* **1996**, *77*, 3865.

[34] A. V. Akimov, O. V. Prezhdo, "Large-scale computations in chemistry: a bird's eye view of a vibrant field", *Chem. Rev.* **2015**, *115*, 5797–5890.

[35] K. Raghavachari, G. W. Trucks, J. A. Pople, M. Head-Gordon, "A fifth-order perturbation comparison of electron correlation theories", *Chem. Phys. Lett.* **1989**, *157*, 479–483.

[36] M. Head-Gordon, "Quantum chemistry and molecular processes", *J. Phys. Chem.* **1996**, *100*, 13213–13225.

[37] L. E. Ratcliff, S. Mohr, G. Huhs, et al., "Challenges in large scale quantum mechanical calculations", *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2017**, *7*, e1290.

[38] D. J. Cole, N. D. Hine, "Applications of large-scale density functional theory in biology", *J. Phys. Condens. Matter* **2016**, *28*, 393001.

[39] R. J. Harrison, R. Shepard, "Ab initio molecular electronic structure on parallel computers", *Annu. Rev. Phys. Chem.* **1994**, *45*, 623–658.

[40] W. Yang, "Direct calculation of electron density in density-functional theory", *Phys. Rev. Lett.* **1991**, *66*, 1438.

[41] S. R. Gadre, R. N. Shirsat, A. C. Limaye, "Molecular tailoring approach for simulation of electrostatic properties", *J. Phys. Chem.* **1994**, *98*, 9165–9169.

[42] M. S. Gordon, D. G. Fedorov, S. R. Pruitt, L. V. Slipchenko, "Fragmentation methods: A route to accurate calculations on large systems", *Chem. Rev.* **2012**, *112*, 632–672.

[43] M. A. Collins, R. P. Bettens, "Energy-based molecular fragmentation methods", *Chem. Rev.* **2015**, *115*, 5607–5642.

[44] K. Raghavachari, A. Saha, "Accurate composite and fragment-based quantum chemical models for large molecules", *Chem. Rev.* **2015**, *115*, 5643–5677.

[45] R. Descartes, *Discourse on the Method of Rightly Conducting one's Reason and Seeking Truth in the Sciences*, in the version by Jonathan Bennett, **1637**, p. 9, `www.earlymoderntexts.com`, accessed on 25/03/2021.

[46] W. Yang, "Direct calculation of electron density in density-functional theory: Implementation for benzene and a tetrapeptide", *Phys. Rev. A* **1991**, *44*, 7823.

[47] J. P. Lu, W. Yang, "Shape of large single- and multiple-shell fullerenes", *Phys. Rev. B* **1994**, *49*, 11421–11424.

[48] W. Yang, T.-S. Lee, "A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules", *J. Chem. Phys.* **1995**, *103*, 5674–5678.

[49] S. L. Dixon, K. M. Merz Jr, "Semiempirical molecular orbital calculations with linear system size scaling", *J. Chem. Phys.* **1996**, *104*, 6643–6649.

[50] S. L. Dixon, K. M. Merz Jr, "Fast, accurate semiempirical molecular orbital calculations for macromolecules", *J. Chem. Phys.* **1997**, *107*, 879–893.

[51] X. He, K. M. Merz Jr, "Divide and conquer Hartree- Fock calculations on proteins", *J. Chem. Theory Comput.* **2010**, *6*, 405–411.

[52] W. Li, S. Li, "Divide-and-conquer local correlation approach to the correlation energy of large molecules", *J. Chem. Phys.* **2004**, *121*, 6649–6657.

[53] M. Kobayashi, T. Akama, H. Nakai, "Second-order Møller-Plesset perturbation energy obtained from divide-and-conquer Hartree-Fock density matrix", *J. Chem. Phys.* **2006**, *125*, 204106.

[54] M. Kobayashi, H. Nakai, "Extension of linear-scaling divide-and-conquer-based correlation method to coupled cluster theory with singles and doubles excitations", *J. Chem. Phys.* **2008**, *129*, 044103.

[55] J. Khandogin, D. M. York, "Quantum descriptors for biological macromolecules from linear-scaling electronic structure methods", *Proteins: Struct. Funct. Bioinf.* **2004**, *56*, 724–737.

[56] T.-S. Lee, D. M. York, W. Yang, "Linear-scaling semiempirical quantum calculations for macromolecules", *J. Chem. Phys.* **1996**, *105*, 2744–2750.

[57] K. M. Merz Jr, "Using quantum mechanical approaches to study biological systems", *Acc. Chem. Res.* **2014**, *47*, 2804–2811.

[58] K. Babu, S. R. Gadre, "Ab initio quality one-electron properties of large molecules: Development and testing of molecular tailoring approach", *J. Comput. Chem.* **2003**, *24*, 484–495.

[59] K. Babu, V. Ganesh, S. R. Gadre, N. E. Ghermani, "Tailoring approach for exploring electron densities and electrostatic potentials of molecular crystals", *Theor. Chem. Acc.* **2004**, *111*, 255–263.

[60] V. Ganesh, K. D. Rameshwar, P. Balanarayan, S. R. Gadre, "Molecular tailoring approach for geometry optimization of large molecules: Energy evaluation and parallelization strategies", *J. Chem. Phys.* **2006**, *125*, 104109.

[61] S. R. Gadre, V. Ganesh, "Molecular tailoring approach: Towards pc-based *ab initio* treatment of large molecules", *J. Theor. Comput. Chem.* **2006**, *5*, 835–855.

[62] N. Sahu, S. R. Gadre, "Molecular tailoring approach: a route for *ab initio* treatment of large clusters", *Acc. Chem. Res.* **2014**, *47*, 2739–2747.

[63] M. Isegawa, B. Wang, D. G. Truhlar, "Electrostatically embedded molecular tailoring approach and validation for peptides", *J. Chem. Theory Comput.* **2013**, *9*, 1381–1393.

[64] A. P. Rahalkar, M. Katouda, S. R. Gadre, S. Nagase, "Molecular tailoring approach in conjunction with MP2 and Ri-MP2 codes: A comparison with fragment molecular orbital method", *J. Comput. Chem.* **2010**, *31*, 2405–2418.

[65] L. Huang, L. Massa, J. Karle, "Kernel energy method illustrated with peptides", *Int. J. Quantum Chem.* **2005**, *103*, 808–817.

[66] L. Huang, L. Massa, J. Karle, "Kernel energy method: Basis functions and quantum methods", *Int. J. Quantum Chem.* **2006**, *106*, 447–457.

[67] L. Huang, L. Massa, J. Karle, "The kernel energy method of quantum mechanical approximation carried to fourth-order terms", *Proc. Natl. Acad. Sci. U.S.A* **2008**, *105*, 1849–1854.

[68] S. N. Weiss, L. Huang, L. Massa, "A generalized higher order kernel energy approximation method", *J. Comput. Chem.* **2010**, *31*, 2889–2899.

[69] L. Huang, L. Massa, J. Karle, "Kernel energy method: Application to insulin", *Proc. Natl. Acad. Sci.* **2005**, *102*, 12690–12693.

[70] L. Huang, L. Massa, J. Karle, "Kernel energy method applied to vesicular stomatitis virus nucleoprotein", *Proc. Natl. Acad. Sci.* **2009**, *106*, 1731–1736.

[71] L. Huang, M. Krupkin, A. Bashan, A. Yonath, L. Massa, "Protoribosome by quantum kernel energy method", *Proc. Natl. Acad. Sci.* **2013**, *110*, 14900–14905.

[72] L. Huang, L. Massa, J. Karle, "Kernel energy method: Application to DNA", *Biochemistry* **2005**, *44*, 16747–16752.

[73] L. Huang, L. Massa, J. Karle, "The kernel energy method: Application to a tRNA", *Proc. Natl. Acad. Sci.* **2006**, *103*, 1233–1237.

[74] L. Huang, H. J. Bohorquez, C. F. Matta, L. Massa, "The kernel energy method: application to graphene and extended aromatics", *Int. J. Quantum Chem.* **2011**, *111*, 4150–4157.

[75] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, M. Uebayasi, "Fragment molecular orbital method: an approximate computational method for large molecules", *Chem. Phys. Lett.* **1999**, *313*, 701–706.

[76] D. G. Fedorov, K. Kitaura in *Modern Methods for Theoretical Physical Chemistry of Biopolymers*, (Eds.: E. B. Starikov, J. P. Lewis, S. Tanaka), Elsevier Science, Amsterdam, **2006**, pp. 3–38.

[77] D. G. Fedorov, K. Kitaura, "Extending the power of quantum chemistry to large systems with the fragment molecular orbital method", *J. Phys. Chem. A* **2007**, *111*, 6904–6914.

[78] D. G. Fedorov, T. Nagata, K. Kitaura, "Exploring chemistry with the fragment molecular orbital method", *Phys. Chem. Chem. Phys.* **2012**, *14*, 7562–7577.

[79] S. Tanaka, Y. Mochizuki, Y. Komeiji, Y. Okiyama, K. Fukuzawa, "Electron-correlated fragment-molecular-orbital calculations for biomolecular and nano systems", *Phys. Chem. Chem. Phys.* **2014**, *16*, 10310–10344.

[80] D. G. Fedorov, "The fragment molecular orbital method: theoretical development, implementation in GAMESS, and applications", *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7*, e1322.

[81] D. Takaya, C. Watanabe, S. Nagase, et al., "FMODB: The World's First Database of Quantum Mechanical Calculations for Biomacromolecules Based on the Fragment Molecular Orbital Method", *J. Chem. Inf. Model.* **2021**, *61*, 777–794.

[82] D. G. Fedorov, T. Ishida, K. Kitaura, "Multilayer formulation of the fragment molecular orbital method (FMO)", *J. Phys. Chem. A* **2005**, *109*, 2638–2646.

[83] Y. Nishimoto, D. G. Fedorov, S. Irle, "Density-functional tight-binding combined with the fragment molecular orbital method", *J. Chem. Theory Comput.* **2014**, *10*, 4801–4812.

[84] D. W. Zhang, J. Z. H. Zhang, "Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein–molecule interaction energy", *J. Chem. Phys.* **2003**, *119*, 3599–3605.

[85] A. M. Gao, D. W. Zhang, J. Z. Zhang, Y. Zhang, "An efficient linear scaling method for ab initio calculation of electron density of proteins", *Chem. Phys. Lett.* **2004**, *394*, 293–297.

[86] X. Chen, Y. Zhang, J. Z. Zhang, "An efficient approach for ab initio energy calculation of biopolymers", *J. Chem. Phys.* **2005**, *122*, 184105.

[87] S. Li, W. Li, T. Fang, "An efficient fragment-based approach for predicting the ground-state energies and structures of large molecules", *J. Am. Chem. Soc.* **2005**, *127*, 7215–7226.

[88] J. Bergmann, M. Davidson, E. Oksanen, U. Ryde, D. Jayatilaka, "fragHAR: towards ab initio quantum-crystallographic X-ray structure refinement for polypeptides and proteins", *IUCrJ* **2020**, *7*, 158–165.

[89] P. G. Mezey, "Shape analysis of macromolecular electron densities", *Structural Chem.* **1995**, *6*, 261–270.

[90] P. G. Mezey in *Advances in quantum chemistry, Vol. 27*, Elsevier, **1996**, pp. 163–222.

[91] P. G. Mezey in *Linear-Scaling Techniques in Computational Chemistry and Physics*, Springer, **2011**, pp. 129–146.

[92] P. G. Mezey, "Fuzzy electron density fragments in macromolecular quantum chemistry, combinatorial quantum chemistry, functional group analysis, and shape–activity relations", *Acc. Chem. Res.* **2014**, *47*, 2821–2827.

[93] P. D. Walker, P. G. Mezey, "Molecular electron density Lego approach to molecule building", *J. Am. Chem. Soc.* **1993**, *115*, 12423–12430.

[94] P. D. Walker, P. G. Mezey, "Toward similarity measures for macromolecular bodies: MEDLA test calculations for substituted benzene systems", *J. Comput. Chem.* **1995**, *16*, 1238–1249.

[95] P. G. Mezey, "Macromolecular density matrices and electron densities with adjustable nuclear geometries", *J. Math. Chem.* **1995**, *18*, 141–168.

[96] P. D. Walker, P. G. Mezey, "*Ab initio* quality electron densities for proteins: A MEDLA approach", *J. Am. Chem. Soc.* **1994**, *116*, 12022–12032.

[97] P. D. Walker, P. G. Mezey, "Realistic, detailed images of proteins and tertiary structure elements: ab initio quality electron density calculations for bovine insulin", *Can. J. Chem.* **1994**, *72*, 2531–2536.

[98] P. D. Walker, P. G. Mezey, "A new computational microscope for molecules: High resolution MEDLA images of taxol and HIV-1 protease, using additive electron density fragmentation principles and fuzzy set methods", *J. Math. Chem.* **1995**, *17*, 203–234.

[99] T. E. Exner, P. G. Mezey, "*Ab initio*-quality electrostatic potentials for proteins: An application of the ADMA approach", *J. Phys. Chem. A* **2002**, *106*, 11791–11800.

[100] T. E. Exner, P. G. Mezey, "Ab initio quality properties for macromolecules using the ADMA approach", *J. Comput. Chem.* **2003**, *24*, 1980–1986.

[101] T. E. Exner, P. G. Mezey, "The field-adapted ADMA approach: Introducing point charges", *J. Phys. Chem. A* **2004**, *108*, 4301–4309.

[102] H. Stoll, G. Wagenblast, H. Preuβ, "On the use of local basis sets for localized molecular orbitals", *Theor. Chim. Acta* **1980**, *57*, 169–178.

[103] A. Fornili, M. Sironi, M. Raimondi, "Determination of extremely localized molecular orbitals and their application to quantum mechanics/molecular mechanics methods and to the study of intramolecular hydrogen bonding", *J. Mol. Struct.: THEOCHEM* **2003**, *632*, 157–172.

[104] M. Sironi, A. Genoni, M. Civera, S. Pieraccini, M. Ghitti, "Extremely localized molecular orbitals: theory and applications", *Theor. Chem. Acc.* **2007**, *117*, 685–698.

[105] B. Meyer, A. Genoni, "Libraries of extremely localized molecular orbitals. 3. construction and preliminary assessment of the new databanks", *J. Phys. Chem. A* **2018**, *122*, 8965–8981.

[106] A. Shurki, B. Braïda, W. Wu in *Complementary Bonding Analysis*, De Gruyter, **2021**, pp. 157–198.

[107] J. C. Tremblay in *Complementary Bonding Analysis*, De Gruyter, **2021**, pp. 113–128.

[108] J. Foster, S. Boys, "Canonical configurational interaction procedure", *Rev. Mod. Phys.* **1960**, *32*, 300–302.

[109] S. F. Boys, "Construction of some molecular orbitals to be approximately invariant for changes from one molecule to another", *Rev. Mod. Phys.* **1960**, *32*, 296.

[110] C. Edmiston, K. Ruedenberg, "Localized atomic and molecular orbitals", *Rev. Mod. Phys.* **1963**, *35*, 457–465.

[111] C. Edmiston, K. Ruedenberg, "Localized atomic and molecular orbitals. II", *J. Chem. Phys.* **1965**, *43*, S97–S116.

[112] J. Pipek, P. G. Mezey, "A fast intrinsic localization procedure applicable for abinitio and semiempirical linear combination of atomic orbital wave functions", *J. Chem. Phys.* **1989**, *90*, 4916–4926.

[113] S. Lehtola, H. Jónsson, "Pipek–Mezey orbital localization using various partial charge estimates", *J. Chem. Theory Comput.* **2014**, *10*, 642–649.

[114] M. Sironi, A. Famulari, "An orthogonal approach to determine extremely localised molecular orbitals", *Theor. Chem. Acc.* **2000**, *103*, 417–422.

[115] M. D. Newton, E. Switkes, W. N. Lipscomb, "Localized Bonds in SCF Wavefunctions for Polyatomic Molecules. III C–H and C–C Bonds", *J. Chem. Phys.* **1970**, *53*, 2645–2657.

[116] M. Levy, W. J. Stevens, H. Shull, S. Hagstrom, "Transferability of electron pairs between H2O and H2O2", *J. Chem. Phys.* **1974**, *61*, 1844–1856.

[117] M Sironi, A. Famulari, M Raimondi, S Chiesa, "The transferability of extremely localized molecular orbitals", *J. Mol. Struct.: THEOCHEM* **2000**, *529*, 47–54.

[118] M. F. Guest, I. J. Bush, H. J. Van Dam, et al., "The GAMESS-UK electronic structure package: algorithms, developments and applications", *Mol. Phys.* **2005**, *103*, 719–747.

[119] B. Meyer, B. Guillot, M. F. Ruiz-Lopez, A. Genoni, "Libraries of extremely localized molecular orbitals. 1. Model molecules approximation and molecular orbitals transferability", *J. Chem. Theory Comput.* **2016**, *12*, 1052–1067.

[120] E. Burresi, M. Sironi, "Determination of extremely localized molecular orbitals in the framework of density functional theory", *Theor. Chem. Acc.* **2004**, *112*, 247–253.

[121] A. Genoni, "X-ray constrained extremely localized molecular orbitals: theory and critical assessment of the new technique", *J. Chem. Theory Comput.* **2013**, *9*, 3004–3019.

[122] D. Jayatilaka, "Wave function for beryllium from X-ray diffraction data", *Phys. Rev. Lett.* **1998**, *80*, 798.

[123] D. Jayatilaka, D. J. Grimwood, "Wavefunctions derived from experiment. I. Motivation and theory", *Acta Crystallogr. A* **2001**, *57*, 76–86.

[124] D. J. Grimwood, D. Jayatilaka, "Wavefunctions derived from experiment. II. A wavefunction for oxalic acid dihydrate", *Acta Crystallogr. A* **2001**, *57*, 87–100.

[125] I. Bytheway, D. J. Grimwood, D. Jayatilaka, "Wavefunctions derived from experiment. III. Topological analysis of crystal fragments", *Acta Crystallogr. A* **2002**, *58*, 232–243.

[126] I. Bytheway, D. J. Grimwood, B. N. Figgis, G. S. Chandler, D. Jayatilaka, "Wavefunctions derived from experiment. IV. Investigation of the crystal environment of ammonia", *Acta Crystallogr. A* **2002**, *58*, 244–251.

[127] D. J. Grimwood, I. Bytheway, D. Jayatilaka, "Wave functions derived from experiment. V. Investigation of electron densities, electrostatic potentials, and electron localization functions for noncentrosymmetric crystals", *J. Comput. Chem.* **2003**, *24*, 470–483.

[128] M. Hudák, D. Jayatilaka, L. Perašínová, et al., "X-ray constrained unrestricted Hartree–Fock and Douglas–Kroll–Hess wavefunctions", *Acta Crystallogr. A* **2010**, *66*, 78–92.

[129] D. Jayatilaka in *Modern Charge-Density Analysis*, (Eds.: C. Gatti, P. Macchi), Springer, Dordrecht, **2012**, pp. 213–257.

[130] D. M. Philipp, R. A. Friesner, "Mixed ab initio QM/MM modeling using frozen orbitals and tests with alanine dipeptide and tetrapeptide", *J. Comput. Chem.* **1999**, *20*, 1468–1494.

[131] A. Genoni, M. Sironi, "A novel approach to relax extremely localized molecular orbitals: the extremely localized molecular orbital–valence bond method", *Theor. Chem. Acc.* **2004**, *112*, 254–262.

[132] A. Genoni, A. Fornili, M. Sironi, "Optimal virtual orbitals to relax wave functions built up with transferred extremely localized molecular orbitals", *J. Comput. Chem.* **2005**, *26*, 827–835.

[133] A. Fornili, Y. Moreau, M. Sironi, X. Assfeld, "On the suitability of strictly localized orbitals for hybrid QM/MM calculations", *J. Comput. Chem.* **2006**, *27*, 515–523.

[134] B. Meyer, B. Guillot, M. F. Ruiz-Lopez, C. Jelsch, A. Genoni, "Libraries of extremely localized molecular orbitals. 2. Comparison with the pseudoatoms transferability", *J. Chem. Theory Comput.* **2016**, *12*, 1068–1081.

[135] M. Sironi, M. Ghitti, A. Genoni, G. Saladino, S. Pieraccini, "DENPOL: A new program to determine electron densities of polypeptides using extremely localized molecular orbitals", *J. Mol. Struct.: THEOCHEM* **2009**, *898*, 8–16.

[136] R. Ditchfield, W. J. Hehre, J. A. Pople, "Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules", *J. Chem. Phys.* **1971**, *54*, 724–728.

[137] W. J. Hehre, R. Ditchfield, J. A. Pople, "Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules", *J. Chem. Phys.* **1972**, *56*, 2257–2261.

[138] M. S. Gordon, J. S. Binkley, J. A. Pople, W. J. Pietro, W. J. Hehre, "Self-consistent molecular-orbital methods. 22. Small split-valence basis sets for second-row elements", *J. Am. Chem. Soc.* **1982**, *104*, 2797–2803.

[139] M. M. Francl, W. J. Pietro, W. J. Hehre, et al., "Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements", *J. Chem. Phys.* **1982**, *77*, 3654–3665.

[140] V. A. Rassolov, M. A. Ratner, J. A. Pople, P. C. Redfern, L. A. Curtiss, "6-31G* basis set for third-row atoms", *J. Comput. Chem.* **2001**, *22*, 976–984.

[141] R. Krishnan, J. S. Binkley, R. Seeger, J. A. Pople, "Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions", *J. Chem. Phys.* **1980**, *72*, 650–654.

[142] A. D. McLean, G. S. Chandler, "Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11-18", *J. Chem. Phys.* **1980**, *72*, 5639–5648.

[143] L. A. Curtiss, M. P. McGrath, J.-P. Blaudeau, et al., "Extension of Gaussian-2 theory to molecules containing third-row atoms Ga-Kr", *J. Chem. Phys.* **1995**, *103*, 6104–6113.

[144] P. C. Hariharan, J. A. Pople, "The influence of polarization functions on molecular orbital hydrogenation energies", *Theor. Chim. Acta* **1973**, *28*, 213–222.

[145] T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen", *J. Chem. Phys.* **1989**, *90*, 1007–1023.

[146] A. K. Wilson, D. E. Woon, K. A. Peterson, T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. IX. The atoms gallium through krypton", *J. Chem. Phys.* **1999**, *110*, 7667–7676.

[147] D. E. Woon, T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon", *J. Chem. Phys.* **1993**, *98*, 1358–1371.

[148] D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, et al., *Amber 2018*, University of California San Francisco, San Francisco, CA, **2018**.

[149] L. A. Malaspina, E. K. Wieduwilt, J. Bergmann, et al., "Fast and accurate quantum crystallography: from small to large, from light to heavy", *J. Phys. Chem. Lett.* **2019**, *10*, 6973–6982.

[150] M. J. Frisch, G. W. Trucks, H. B. Schlegel, et al., Gaussian 09 Revision D.01, **2009**.

[151] Q. Sun, G. K.-L. Chan, "Quantum embedding theories", *Acc. Chem. Res.* **2016**, *49*, 2705–2712.

[152] H. M. Senn, W. Thiel, "QM/MM methods for biomolecular systems", *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

[153] A. Warshel, M. Levitt, "Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme", *J. Mol. Biol.* **1976**, *103*, 227–249.

[154] B. Mennucci, "Continuum solvation models: What else can we learn from them?", *J. Phys. Chem. Lett.* **2010**, *1*, 1666–1674.

[155] B. Mennucci, "Polarizable continuum model", *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 386–404.

[156] D. Rinaldi, J.-L. Rivail, "Polarisabilites moléculaires et effet diélectrique de milieu à l'état liquide. Étude théorique de la molécule d'eau et de ses diméres", *Theor. Chim. Acta* **1973**, *32*, 57–70.

[157] J.-L. Rivail, D. Rinaldi, "A quantum chemical approach to dielectric solvent effects in molecular liquids", *Chem. Phys.* **1976**, *18*, 233–242.

[158] D. Rinaldi, M. F. Ruiz-Lopez, J.-L. Rivail, "Ab initio SCF calculations on electrostatically solvated molecules using a deformable three axes ellipsoidal cavity", *J. Chem. Phys.* **1983**, *78*, 834–838.

[159] L. O. Jones, M. A. Mosquera, G. C. Schatz, M. A. Ratner, "Embedding methods for quantum chemistry: applications from materials to life sciences", *J. Am. Chem. Soc.* **2020**, *142*, 3281–3295.

[160] Nobel Media AB 2021, The Nobel Prize in Chemistry 2013, `https://www.nobelprize.org/prizes/chemistry/2013/summary/`, accessed on 16/04/2021.

[161] M. Karplus, "Development of multiscale models for complex chemical systems: from H+ H2 to biomolecules (Nobel lecture)", *Angew. Chem. Int. Ed.* **2014**, *53*, 9992–10005.

[162] M. Levitt, "Birth and future of multiscale modeling for macromolecular systems (Nobel Lecture)", *Angew. Chem. Int. Ed.* **2014**, *53*, 10006–10018.

[163] A. Warshel, "Multiscale modeling of biological functions: from enzymes to molecular machines (Nobel Lecture)", *Angew. Chem. Int. Ed.* **2014**, *53*, 10020–10031.

[164] A. D. MacKerell Jr, "Empirical force fields for biological macromolecules: overview and issues", *J. Comput. Chem.* **2004**, *25*, 1584–1604.

[165] U Ryde, "QM/MM calculations on proteins", *Methods Enzymol.* **2016**, *577*, 119–158.

[166] L. Cao, U. Ryde, "On the difference between additive and subtractive QM/MM calculations", *Front. Chem.* **2018**, *6*, 89.

[167] V. Théry, D. Rinaldi, J.-L. Rivail, B. Maigret, G. G. Ferenczy, "Quantum mechanical computations on very large molecular systems: The local self-consistent field method", *J. Comput. Chem.* **1994**, *15*, 269–282.

[168] G. Monard, M. Loos, V. Théry, K. Baka, J.-L. Rivail, "Hybrid classical quantum force field for modeling very large molecules", *Int. J. Quantum Chem.* **1996**, *58*, 153–159.

[169] X. Assfeld, J.-L. Rivail, "Quantum chemical computations on parts of large molecules: the ab initio local self consistent field method", *Chem. Phys. Lett.* **1996**, *263*, 100–106.

[170] N. Ferré, X. Assfeld, J.-L. Rivail, "Specific force field parameters determination for the hybrid ab initio QM/MM LSCF method", *J. Comput. Chem.* **2002**, *23*, 610–624.

[171] M. Svensson, S. Humbel, R. D. Froese, et al., "ONIOM: a multilayered integrated MO+ MM method for geometry optimizations and single point energy predictions. A test for Diels- Alder reactions and Pt (P (t-Bu) 3) 2+ H2 oxidative addition", *J. Phys. Chem.* **1996**, *100*, 19357–19363.

[172] L. W. Chung, W. Sameera, R. Ramozzi, et al., "The ONIOM method and its applications", *Chem. Rev.* **2015**, *115*, 5678–5796.

[173] A Jack, M Levitt, "Refinement of large structures by simultaneous minimization of energy and R factor", *Acta Crystallogr. A* **1978**, *34*, 931–935.

[174] J. Bergmann, E. Oksanen, U. Ryde, "Combining crystallography with quantum mechanics", *Curr. Opin. Struct. Biol.* **2022**, *72*, 18–26.

[175] A. Genoni, L. Bučinskỳ, N. Claiser, et al., "Quantum crystallography: Current developments and future perspectives", *Chem. Eur. J.* **2018**, *24*, 10881–10905.

[176] U. Ryde, L. Olsen, K. Nilsson, "Quantum chemical geometry optimizations in proteins using crystallographic raw data", *J. Comput. Chem.* **2002**, *23*, 1058–1070.

[177] U. Ryde, K. Nilsson, "Quantum chemistry can locally improve protein crystal structures", *J. Am. Chem. Soc.* **2003**, *125*, 14232–14233.

[178] N. Yu, H. P. Yennawar, K. M. Merz, "Refinement of protein crystal structures using energy restraints derived from linear-scaling quantum mechanics", *Acta Crystallogr. D* **2005**, *61*, 322–332.

[179] F. R. Manby, M. Stella, J. D. Goodpaster, T. F. Miller III, "A simple, exact density-functional-theory embedding scheme", *J. Chem. Theory Comput.* **2012**, *8*, 2564–2568.

[180] T. A. Wesołowski, A. Warshel, "Frozen density functional approach for ab initio calculations of solvated molecules", *J. Phys. Chem.* **1993**, *97*, 8050–8053.

[181] T. A. Wesołowski, "Embedding a multideterminantal wave function in an orbital-free environment", *Phys. Rev. A* **2008**, *77*, 012504.

[182] T. A. Wesołowski, S. Shedge, X. Zhou, "Frozen-density embedding strategy for multilevel simulations of electronic structure", *Chem. Rev.* **2015**, *115*, 5891–5928.

[183] C. R. Jacob, J. Neugebauer, "Subsystem density-functional theory", *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 325–362.

[184] C. R. Jacob, J. Neugebauer, L. Jensen, L. Visscher, "Comparison of frozen-density embedding and discrete reaction field solvent models for molecular properties", *Phys. Chem. Chem. Phys.* **2006**, *8*, 2349–2359.

[185] T. A. Wesoołowski, J. Weber, "Kohn-Sham equations with constrained electron density: an iterative evaluation of the ground-state electron density of interacting molecules", *Chem. Phys. Lett.* **1996**, *248*, 71–76.

[186] M. E. Casida, T. A. Wesołowski, "Generalization of the Kohn–Sham equations with constrained electron density formalism and its time-dependent response theory formulation", *Int. J. Quantum Chem.* **2004**, *96*, 577–588.

[187] J. Neugebauer, "Couplings between electronic transitions in a subsystem formulation of time-dependent density functional theory", *J. Chem. Phys* **2007**, *126*, 134116.

[188] J. Neugebauer, "Photophysical properties of natural light-harvesting complexes studied by subsystem density functional theory", *J. Phys. Chem. B* **2008**, *112*, 2207–2217.

[189] J. Neugebauer, M. J. Louwerse, E. J. Baerends, T. A. Wesolowski, "The merits of the frozen-density embedding scheme to model solvatochromic shifts", *J. Chem. Phys.* **2005**, *122*, 094115.

[190] J. Neugebauer, C. R. Jacob, T. A. Wesolowski, E. J. Baerends, "An explicit quantum chemical method for modeling large solvation shells applied to aminocoumarin C151", *J. Phys. Chem. A* **2005**, *109*, 7805–7814.

[191] S. J. Lee, M. Welborn, F. R. Manby, T. F. Miller III, "Projection-based wavefunction-in-DFT embedding", *Acc. Chem. Res.* **2019**, *52*, 1359–1368.

[192] D. V. Chulhai, J. D. Goodpaster, "Improved accuracy and efficiency in quantum embedding through absolute localization", *J. Chem. Theory Comput.* **2017**, *13*, 1503–1508.

[193] D. V. Chulhai, J. D. Goodpaster, "Projection-based correlated wave function in density functional theory embedding for periodic systems", *J. Chem. Theory Comput.* **2018**, *14*, 1928–1942.

[194] F Libisch, M Marsman, J Burgdörfer, G Kresse, "Embedding for bulk systems using localized atomic orbitals", *J. Chem. Phys.* **2017**, *147*, 034110.

[195] S. J. Bennie, M. W. van der Kamp, R. C. Pennifold, et al., "A projector-embedding approach for multiscale coupled-cluster calculations applied to citrate synthase", *J. Chem. Theory Comput.* **2016**, *12*, 2689–2697.

[196] K. E. Ranaghan, D. Shchepanovska, S. J. Bennie, et al., "Projector-based embedding eliminates density functional dependence for QM/MM calculations of reactions in enzymes and solution", *J. Chem. Theory Comput.* **2019**, *59*, 2063–2078.

[197] G. Marrazzini, T. Giovannini, M. Scavino, et al., "Multilevel density functional theory", *J. Chem. Theory Comput.* **2021**, *17*, 791–803.

[198] A. M. Sanchez de Meras, H. Koch, I. G. Cuesta, L. Boman, "Cholesky decomposition-based definition of atomic subsystems in electronic structure calculations", *J. Chem. Phys.* **2010**, *132*, 204105.

[199] S. Sæther, T. Kjærgaard, H. Koch, I.-M. Høyvik, "Density-based multilevel Hartree–Fock model", *J. Chem. Theory Comput.* **2017**, *13*, 5282–5290.

[200] I.-M. Høyvik, "Convergence acceleration for the multilevel Hartree–Fock model", *Mol. Phys.* **2020**, *118*, 1626929.

[201] R. H. Myhre, A. M. Sánchez de Merás, H. Koch, "Multi-level coupled cluster theory", *J. Chem. Phys.* **2014**, *141*, 224105.

[202] S. D. Folkestad, H. Koch, "Multilevel CC2 and CCSD Methods with Correlated Natural Transition Orbitals", *J. Chem. Theory Comput.* **2019**, *16*, 179–189.

[203] R. H. Myhre, H. Koch, "The multilevel CC3 coupled cluster model", *J. Chem. Phys.* **2016**, *145*, 044111.

[204] L. Goletto, T. Giovannini, S. D. Folkestad, H. Koch, "Combining multilevel Hartree–Fock and multilevel coupled cluster approaches with molecular mechanics: a study of electronic excitations in solutions", *Phys. Chem. Chem. Phys.* **2021**, *23*, 4413–4425.

[205] G. Macetti, A. Genoni, "Quantum mechanics/extremely localized molecular orbital method: A fully quantum mechanical embedding approach for macromolecules", *J. Phys. Chem. A* **2019**, *123*, 9420–9428.

[206] G. Macetti, E. K. Wieduwilt, X. Assfeld, A. Genoni, "Localized molecular orbital-based embedding scheme for correlated methods", *J. Chem. Theory Comput.* **2020**, *16*, 3578–3596.

[207] G. Macetti, E. K. Wieduwilt, A. Genoni, "QM/ELMO: A multi-purpose fully quantum mechanical embedding scheme based on extremely localized molecular orbitals", *J. Phys. Chem. A* **2021**, *125*, 2709–2726.

[208] G. Macetti, A. Genoni, "Quantum Mechanics/Extremely Localized Molecular Orbital Embedding Strategy for Excited States: Coupling to Time-Dependent Density Functional Theory and Equation-of-Motion Coupled Cluster", *J. Chem. Theory Comput.* **2020**, *16*, 7490–7506.

[209] G. Macetti, A. Genoni, "Initial Maximum Overlap Method for Large Systems by the Quantum Mechanics/Extremely Localized Molecular Orbital Embedding Technique", *J. Chem. Theory Comput.* **2021**, *17*, 4169–4182.

[210] G. Macetti, A. Genoni, "Three-Layer Multiscale Approach Based on Extremely Localized Molecular Orbitals to Investigate Enzyme Reactions", *J. Phys. Chem. A* **2021**, *125*, 6013–6027.

[211] J. C. Brooks-Bartlett, E. F. Garman, "The nobel science: one hundred years of crystallography", *Interdiscip. Sci. Rev.* **2015**, *40*, 244–264.

[212] M. Jaskolski, Z. Dauter, A. Wlodawer, "A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits", *FEBS J.* **2014**, *281*, 3985–4009.

[213] G. M. Sheldrick, "A short history of SHELX", *Acta Crystallogr. A* **2008**, *64*, 112–122.

[214] J. S. Richardson, "The Anatomy and Taxonomy of Protein Structure (and updates)", *Advances in Protein Chemistry web version* **1981 & 2000-2007**, *34*, (Eds.: C. B. Anfinsen, J. T. Edsall, F. M. Richards), 167–339, `http://kinemage.biochem.duke.edu/teaching/anatax/`.

[215] J. D. Watson, F. H. C. Crick, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid", *Nature* **1953**, *171*, 737–738.

[216] M. H. F. Wilkins, A. R. Stokes, H. R. Wilson, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid", *Nature* **1953**, *171*, 738–740.

[217] R. E. Franklin, R. G. Gosling, "Molecular configuration in sodium thymonucleate", *Nature* **1953**, *171*, 740–741.

[218] J. C. Kendrew, G Bodo, H. M. Dintzis, et al., "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis", *Nature* **1958**, *181*, 662–666.

[219] H. M. Berman, J. Westbrook, Z. Feng, et al., "The protein data bank", *Nucleic Acids Res.* **2000**, *28*, 235–242.

[220] A. H. Compton, "The distribution of the electrons in atoms", *Nature* **1915**, *95*, 343–344.

[221] E. K. Wieduwilt, G. Macetti, L. A. Malaspina, et al., "Post-Hartree-Fock methods for Hirshfeld atom refinement: are they necessary? Investigation of a strongly hydrogen-bonded molecular crystal", *J. Mol. Struct.* **2020**, *1209*, 127934.

[222] J. M. Word, S. C. Lovell, T. H. LaBean, et al., "Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms", *J. Mol. Biol.* **1999**, *285*, 1711–1733.

[223] T. Steiner, "The hydrogen bond in the solid state", *Angew. Chem.* **2002**, *41*, 48–76.

[224] D. Braga, F. Grepioni, K. Biradha, G. R. Desiraju, "Agostic interactions in organometallic compounds. A Cambridge Structural Database study", *J. Chem. Soc. Dalton Trans.* **1996**, 3925–3930.

[225] S. Grabowsky, A. Genoni, H.-B. Bürgi, "Quantum crystallography", *Chem. Sci.* **2017**, *8*, 4159–4176.

[226] P. Macchi, "The connubium between crystallography and quantum mechanics", *Crystallogr. Rev.* **2020**, *26*, 209–268.

[227] D. Jayatilaka, B. Dittrich, "X-ray structure refinement using aspherical atomic density functions obtained from quantum-mechanical calculations", *Acta Crystallogr. A* **2008**, *64*, 383–393.

[228] S. C. Capelli, H.-B. Bürgi, B. Dittrich, S. Grabowsky, D. Jayatilaka, "Hirshfeld atom refinement", *IUCrJ* **2014**, *1*, 361–379.

[229] M. Woińska, S. Grabowsky, P. M. Dominiak, K. Woźniak, D. Jayatilaka, "Hydrogen atoms can be located accurately and precisely by x-ray crystallography", *Sci. Adv.* **2016**, *2*, e1600192.

[230] H. Wondratschek in *International tables for crystallography*, *Vol. A*, (Ed.: M. I. Aroyo), Wiley Online Library, **2006**, pp. 720–725.

[231] W. Massa, *Kristallstrukturbestimmung*, *Vol. 8*, Springer, **2015**.

[232] W. Friedrich, P. Knipping, M. Laue in *Sitzungsberichte der mathematisch-physikalischen Klasse*, Verlag der Kgl. Bayer. Akad. der Wiss., **1912**, pp. 303–322.

[233] M. Eckert, "Max von Laue and the discovery of X-ray diffraction in 1912", *Ann. Phys. (Berl.)* **2012**, *524*, A83–A85.

[234] M. von Laue, Nobel Lecture, `https://www.nobelprize.org/prizes/physics/1914/laue/lecture`, accessed on 14/05/2021.

[235] W. H. Bragg, W. L. Bragg, "The reflection of X-rays by crystals", *Proc. Royal Soc. London Ser. Containing Papers Math. Phys. Charact.* **1913**, *88*, 428–438.

[236] Philip Coppens, *X-ray charge densities and chemical bonding*, Oxford Univ. Press, Oxford, **1997**, pp. 1–21.

[237] C. Gatti, P. Macchi in *Modern Charge-Density Analysis*, (Eds.: C. Gatti, P. Macchi), Springer, Dordrecht, **2012**, pp. 1–78.

[238] P. Luger in *Modern X-Ray Analysis on Single Crystals*, De Gruyter, **2014**, pp. 14–42.

[239] P. J. Brown, A. G. Fox, E. N. Maslen, M. A. O'Keefe, B. T. M Willis in *International tables for crystallography, Vol. C*, (Ed.: E. Prince), Wiley Online Library, **2006**, pp. 554–595.

[240] A. Goeta, J. Howard, "Low temperature single crystal X-ray diffraction: advantages, instrumentation and applications", *Chem. Soc. Rev.* **2004**, *33*, 490–500.

[241] G. L. Taylor, "Introduction to phasing", *Acta Crystallogr. D* **2010**, *66*, 325–338.

[242] P. Müller, "Practical suggestions for better crystal structures", *Crystallogr. Rev.* **2009**, *15*, 57–83.

[243] A. Wlodawer, W. Minor, Z. Dauter, M. Jaskolski, "Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination", *FEBS J.* **2013**, *280*, 5705–5736.

[244] R. Destro, R. E. Marsh, R. Bianchi, "A low-temperature (23 K) study of L-alanine", *J. Phys. Chem.* **1988**, *92*, 966–973.

[245] G. M. Sheldrick, "Phase annealing in SHELX-90: direct methods for larger structures", *Acta Crystallogr. A* **1990**, *46*, 467–473.

[246] R. J. Morris, G. Bricogne, "Sheldrick's 1.2 Å rule and beyond", *Acta Crystallogr. D* **2003**, *59*, 615–617.

[247] H. Ahmed, M. Blakeley, M Cianci, et al., "The determination of protonation states in proteins", *Acta Crystallogr. D* **2007**, *63*, 906–922.

[248] A. L. Spek, "checkCIF validation ALERTS: what they mean and how to respond", *Acta Crystallogr. E* **2020**, *76*, 1–11.

[249] K. Meindl, J. Henn, "Foundations of residual-density analysis", *Acta Crystallogr. A* **2008**, *64*, 404–418.

[250] L. A. Malaspina, A. J. Edwards, M. Woińska, et al., "Predicting the Position of the Hydrogen Atom in the Short Intramolecular Hydrogen Bond of the Hydrogen Maleate Anion from Geometric Correlations", *Cryst. Growth Des.* **2017**, *17*, 3812–3825.

[251] J. M. Bąk, S. Domagała, C. Hübschle, et al., "Verification of structural and electrostatic properties obtained by the use of different pseudoatom databases", *Acta Crystallogr. A* **2011**, *67*, 141–153.

[252] J Peters, W Jauch, "Single crystal time-of-flight neutron diffraction", *Sci. Prog.* **2002**, *85*, 297–318.

[253] A. J. Edwards, "Neutron diffraction - recent applications to chemical structure determination", *Aust. J. Chem.* **2011**, *64*, 869–872.

[254] F. H. Allen, I. J. Bruno, "Bond lengths in organic and metal-organic compounds revisited: X—H bond lengths from neutron diffraction data", *Acta Crystallogr. B* **2010**, *66*, 380–386.

[255] B. Dittrich, J. Lübben, S. Mebs, et al., "Accurate bond lengths to hydrogen atoms from single-crystal X-ray diffraction by including estimated hydrogen ADPs and comparison to neutron and QM/MM benchmarks", *Chem. Eur. J.* **2017**, *23*, 4605–4614.

[256] V. L. Deringer, V. Hoepfner, R. Dronskowski, "Accurate hydrogen positions in organic crystals: Assessing a quantum-chemical aide", *Cryst. Growth Des.* **2012**, *12*, 1014–1021.

[257] B Dittrich, C. B. Hübschle, M Messerschmidt, et al., "The invariom model and its application: refinement of D, L-serine at different temperatures and resolution", *Acta Crystallogr. A* **2005**, *61*, 314–320.

[258] K. K. Jha, B. Gruza, P. Kumar, M. L. Chodkiewicz, P. M. Dominiak, "TAAM: a reliable and user friendly tool for hydrogen-atom location using routine X-ray diffraction data", *Acta Crystallogr. B* **2020**, *76*.

[259] W. F. Sanjuan-Szklarz, M. Woińska, S. Domagała, et al., "On the accuracy and precision of X-ray and neutron diffraction results as a function of resolution and the electron density model", *IUCrJ* **2020**, *7*.

[260] J. Lübben, C. M. Wandtke, C. B. Hübschle, et al., "Aspherical scattering factors for SHELXL – model, implementation and application", *Acta Crystallogr. A* **2019**, *75*, 50–62.

[261] M. Woińska, D. Jayatilaka, M. A. Spackman, et al., "Hirshfeld atom refinement for modelling strong hydrogen bonds", *Acta Crystallogr. A* **2014**, *70*, 483–498.

[262] M. Woińska, M. L. Chodkiewicz, K. Woźniak, "Towards accurate and precise positions of hydrogen atoms bonded to heavy metal atoms", *Chem. Commun.* **2021**, *57*, 3652–3655.

[263] A. Ø. Madsen, H. O. Sørensen, C. Flensburg, R. F. Stewart, S. Larsen, "Modeling of the nuclear parameters for H atoms in X-ray charge-density studies", *Acta Crystallogr. A* **2004**, *60*, 550–561.

[264] P Coppens, "Comparative X-ray and neutron diffraction study of bonding effects in s-triazine", *Science* **1967**, *158*, 1577–1579.

[265] B. Iversen, F. Larsen, B. Figgis, P. Reynolds, A. Schultz, "Atomic displacement parameters for Ni (ND3) 4 (NO2) 2 from 9 K X-ray and 13 K time-of-flight neutron diffraction data", *Acta Crystallogr. B* **1996**, *52*, 923–931.

[266] P. Munshi, A. Ø. Madsen, M. A. Spackman, S. Larsen, R. Destro, "Estimated H-atom anisotropic displacement parameters: a comparison between different methods and with neutron diffraction results", *Acta Crystallogr. A* **2008**, *64*, 465–475.

[267] A. Ø. Madsen, "SHADE web server for estimation of hydrogen anisotropic displacement parameters", *J. Appl. Cryst.* **2006**, *39*, 757–758.

[268] A. Ø. Madsen, A. A. Hoser, "SHADE3 server: a streamlined approach to estimate H-atom anisotropic displacement parameters using periodic ab initio calculations or experimental information", *J. Appl. Cryst.* **2014**, *47*, 2100–2104.

[269] A. E. Whitten, M. A. Spackman, "Anisotropic displacement parameters for H atoms using an ONIOM approach", *Acta Crystallogr. B* **2006**, *62*, 875–888.

[270] J. Lübben, L. J. Bourhis, B. Dittrich, "Estimating temperature-dependent anisotropic hydrogen displacements with the invariom database and a new segmented rigid-body analysis program", *J. Appl. Cryst.* **2015**, *48*, 1785–1793.

[271] B Dittrich, C. Hübschle, K Pröpper, et al., "The generalized invariom database (GID)", *Acta Crystallogr. B* **2013**, *69*, 91–104.

[272] A. A. Hoser, A. Ø. Madsen, "Dynamic quantum crystallography: lattice-dynamical models refined against diffraction data. I. Theory", *Acta Crystallogr. A* **2016**, *72*, 206–214.

[273] A. A. Hoser, A. Ø. Madsen, "Dynamic quantum crystallography: lattice-dynamical models refined against diffraction data. II. Applications to l-alanine, naphthalene and xylitol", *Acta Crystallogr. A* **2017**, *73*, 102–114.

[274] V. V. Zhurov, E. A. Zhurova, A. I. Stash, A. A. Pinkerton, "Importance of the consideration of anharmonic motion in charge-density studies: a comparison of variable-temperature studies on two explosives, RDX and HMX", *Acta Crystallogr. A* **2011**, *67*, 160–173.

[275] C. Köhler, J. Lübben, L. Krause, et al., "Comparison of different strategies for modelling hydrogen atoms in charge density analyses", *Acta Crystallogr. B* **2019**, *75*, 434–441.

[276] L. A. Malaspina, A. A. Hoser, A. J. Edwards, et al., "Hydrogen atoms in bridging positions from quantum crystallographic refinements: influence of hydrogen atom displacement parameters on geometry and electron density", *Cryst. Eng. Comm* **2020**, *22*, 4778–4789.

[277] P Coppens in *International tables for crystallography, Vol. B*, (Ed.: U. Shmueli), Wiley Online Library, **2010**, pp. 10–23.

[278] V. F. Sears in *International tables for crystallography, Vol. C*, (Ed.: E. Prince), Wiley Online Library, **2006**, pp. 444–454.

[279] P. Luger in *Modern X-Ray Analysis on Single Crystals*, De Gruyter, **2014**, pp. 244–247.

[280] N. K. Hansen, P. Coppens, "Testing aspherical atom refinements on small-molecule data sets", *Acta Crystallogr. A* **1978**, *34*, 909–921.

[281] T. S. Koritsanszky, P. Coppens, "Chemical applications of X-ray charge-density analysis", *Chem. Rev.* **2001**, *101*, 1583–1628.

[282] R. F. Stewart, J. Bentley, B. Goodman, "Generalized x-ray scattering factors in diatomic molecules", *J. Chem. Phys.* **1975**, *63*, 3786–3793.

[283] P. Roversi, F. Merati, R. Destro, M. Barzaghi, "Charge density in crystalline citrinin from X-ray diffraction at 19 K", *Can. J. Chem.* **1996**, *74*, 1145–1161.

[284] R. Destro, F. Merati, "Bond lengths, and beyond", *Acta Crystallogr. B* **1995**, *51*, 559–570.

[285] C. Gatti, E. May, R. Destro, F. Cargnoni, "Fundamental Properties and Nature of CH··O Interactions in Crystals on the Basis of Experimental and Theoretical Charge Densities. The Case of 3,4-Bis(dimethylamino)-3-cyclobutene-1,2-dione (DMACB) Crystal", *J. Phys. Chem. A* **2002**, *106*, 2707–2720.

[286] M. Wanat, M. Malinska, M. J. Gutmann, R. I. Cooper, K. Wozniak, et al., "HAR, TAAM and BODD refinements of model crystal structures using Cu Kα and Mo Kα X-ray diffraction data", *Acta Crystallogr. B* **2021**, *77*, 0–0.

[287] S. Domagała, B. Fournier, D. Liebschner, B. Guillot, C. Jelsch, "An improved experimental databank of transferable multipolar atom models – ELMAM2. Construction details and applications", *Acta Crystallogr. A* **2012**, *68*, 337–351.

[288] K. N. Jarzembska, P. M. Dominiak, "New version of the theoretical databank of transferable aspherical pseudoatoms, UBDB2011 – towards nucleic acid modelling", *Acta Crystallogr. A* **2012**, *68*, 139–147.

[289] C. P. Brock, J. Dunitz, F. Hirshfeld, "Transferability of deformation densities among related molecules: atomic multipole parameters from perylene for improved estimation of molecular vibrations in naphthalene and anthracene", *Acta Crystallogr. B* **1991**, *47*, 789–797.

[290] C. Lecomte, C. Jelsch, B. Guillot, B. Fournier, A. Lagoutte, "Ultrahigh-resolution crystallography and related electron density and electrostatic properties in proteins", *J. Synchrotron Rad.* **2008**, *15*, 202–203.

[291] M. Elias, D. Liebschner, J. Koepke, et al., "Hydrogen atoms in protein structures: high-resolution X-ray diffraction structure of the DFPase", *BMC research notes* **2013**, *6*, 1–7.

[292] K Pröpper, J. Holstein, C. Hübschle, C. Bond, B Dittrich, "Invariom refinement of a new monoclinic solvate of thiostrepton at 0.64 Å resolution", *Acta Crystallogr. D* **2013**, *69*, 1530–1539.

[293] M. Malinska, Z. Dauter, "Transferable aspherical atom model refinement of protein and DNA structures against ultrahigh-resolution X-ray data", *Acta Crystallogr. D* **2016**, *72*, 770–779.

[294] G. M. Sheldrick, "Crystal structure refinement with SHELXL", *Acta Crystallogr. C* **2015**, *71*, 3–8.

[295] F. L. Hirshfeld, "XVII. Spatial partitioning of charge density", *Isr. J. Chem.* **1977**, *16*, 198–201.

[296] F. L. Hirshfeld, "Bonded-atom fragments for describing molecular charge densities", *Theor. Chim. Acta* **1977**, *44*, 129–138.

[297] D. Jayatilaka, D. J. Grimwood in International Conference on Computational Science, Springer, **2003**, pp. 142–151.

[298] M. Fugel, D. Jayatilaka, E. Hupf, et al., "Probing the accuracy and precision of Hirshfeld atom refinement with HARt interfaced with Olex2", *IUCrJ* **2018**, *5*, 32–44.

[299] O. V. Dolomanov, L. J. Bourhis, R. J. Gildea, J. A. Howard, H. Puschmann, "OLEX2: a complete structure solution, refinement and analysis program", *J. Appl. Crystallogr.* **2009**, *42*, 339–341.

[300] L. A. Malaspina, A. Genoni, S. Grabowsky, "lamaGOET: an interface for quantum crystallography", *J. Appl. Cryst.* **2021**, *54*.

[301] F. Neese, "The ORCA program system", *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73–78.

[302] F. Neese, "Software update: the ORCA program system, version 4.0", *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, e1327.

[303] F. Kleemiss, O. V. Dolomanov, M. Bodensteiner, et al., "Accurate crystal structures and chemical properties from NoSpherA2", *Chem. Sci.* **2021**, *12*, 1675–1692.

[304] L. J. Bourhis, O. V. Dolomanov, R. J. Gildea, J. A. Howard, H. Puschmann, "The anatomy of a comprehensive constrained, restrained refinement program for the modern computing environment – Olex2 dissected", *Acta Crystallogr. A* **2015**, *71*, 59–75.

[305] A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior", *Phys. Rev. A* **1988**, *38*, 3098–3100.

[306] C. Lee, W. Yang, R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density", *Phys. Rev. B* **1988**, *37*, 785–789.

[307] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange", *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[308] S. H. Vosko, L. Wilk, M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis", *Can. J. Phys.* **1980**, *58*, 1200–1211.

[309] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, "Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields", *J. Phys. Chem.* **1994**, *98*, 11623–11627.

[310] R. E. Cachau, J. Zhu, M. C. Nicklaus, "The upcoming subatomic resolution revolution", *Curr. Opin. Struct. Biol.* **2019**, *58*, 53–58.

[311] T Petrova, A Podjarny, "Protein crystallography at subatomic resolution", *Rep. Prog. Phys.* **2004**, *67*, 1565.

[312] B. Bax, C.-w. Chung, C. Edge, "Getting the chemistry right: protonation, tautomers and the importance of H atoms in biological chemistry", *Acta Crystallogr. D* **2017**, *73*, 131–140.

[313] S. C. Capelli, H.-B. Bürgi, S. A. Mason, D. Jayatilaka, "Glycyl-L-alanine: a multi-temperature neutron study", *Acta Crystallogr. C* **2014**, *70*, 949–952.

[314] V. Pichon-Pesme, C. Lecomte, R. Wiest, M. Benard, "Modeling fragments for the ab initio determination of electron density in polypeptides. An experimental and theoretical approach to the electron distribution in Leu-enkephalin trihydrate", *J. Am. Chem. Soc.* **1992**, *114*, 2713–2715.

[315] M. R. Sawaya, S. Sambashivan, R. Nelson, et al., "Atomic structures of amyloid cross-β spines reveal varied steric zippers", *Nature* **2007**, *447*, 453–457.

[316] C. Jelsch, M. M. Teeter, V. Lamzin, et al., "Accurate protein crystallography at ultra-high resolution: valence electron distribution in crambin", *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 3171–3176.

[317] A. L. Spek, "PLATON SQUEEZE: a tool for the calculation of the disordered solvent contribution to the calculated structure factors", *Acta Crystallogr. C* **2015**, *71*, 9–18.

[318] M Kohout, "A measure of electron localizability", *Int. J. Quantum Chem.* **2004**, *97*, 651–658.

[319] E. R. Johnson, S. Keinan, P. Mori-Sánchez, et al., "Revealing noncovalent interactions", *J. Am. Chem. Soc.* **2010**, *132*, 6498–6506.

[320] D. Arias-Olivares, E. K. Wieduwilt, J. Contreras-García, A. Genoni, "NCI-ELMO: A new method to quickly and accurately detect noncovalent interactions in biosystems", *J. Chem. Theory Comput.* **2019**, *15*, 6456–6470.

[321] C. Lefebvre, H. Khartabil, J.-C. Boisson, et al., "The independent gradient model: a new approach for probing strong and weak interactions in molecules from wave function calculations", *ChemPhysChem* **2018**, *19*, 724–735.

[322] E. K. Wieduwilt, J.-C. Boisson, G. Terraneo, E. Hénon, A. Genoni, "A Step toward the quantification of noncovalent interactions in large biological systems: The independent gradient model-extremely localized molecular orbital approach", *J. Chem. Inf. Model.* **2021**, *61*, 795–809.

[323] F. Kleemiss, E. K. Wieduwilt, E. Hupf, et al., "Similarities and differences between crystal and enzyme environmental effects on the electron density of drug molecules", *Chem. Eur. J.* **2021**, *27*, 3407.

[324] M. Fugel, F. Kleemiss, L. A. Malaspina, et al., "Investigating the resonance in nitric acid and the nitrate anion based on a modern bonding analysis", *Aust. J. Chem* **2018**, *71*, 227–237.

[325] L. Bučinskỳ, S. Biskupič, D. Jayatilaka, "Study of the picture change error at the 2nd order Douglas Kroll Hess level of theory. Electron and spin density and structure factors of the Bis [bis (methoxycarbimido) aminato] copper (II) complex", *Chem. Phys.* **2012**, *395*, 44–53.

[326] L. Bucinsky, D. Jayatilaka, S. Grabowsky, "Importance of relativistic effects and electron correlation in structure factors and electron density of diphenyl mercury and triphenyl bismuth", *J. Phys. Chem. A* **2016**, *120*, 6650–6669.

[327] L. Bučinskỳ, D. Jayatilaka, S. Grabowsky, "Relativistic quantum crystallography of diphenyl-and dicyanomercury. Theoretical structure factors and Hirshfeld atom refinement", *Acta Crystallogr. A* **2019**, *75*, 705–717.

[328] M. E. Wall, "Quantum crystallographic charge density of urea", *IUCrJ* **2016**, *3*, 237–246.

[329] E. K. Wieduwilt, G. Macetti, A. Genoni, "Climbing Jacob's ladder of structural refinement: Introduction of a localized molecular orbital-based embedding for accurate X-ray determinations of hydrogen atom positions", *J. Phys. Chem. Lett.* **2021**, *12*, 463–471.

[330] M. L. Chodkiewicz, M. Woińska, K. Woźniak, "Hirshfeld atom like refinement with alternative electron density partitions", *IUCrJ* **2020**, *7*.

[331] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, "Development and testing of a general amber force field", *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[332] A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly, "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method", *J. Comput. Chem.* **2000**, *21*, 132–146.

[333] A. Jakalian, D. B. Jack, C. I. Bayly, "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation", *J. Comput. Chem.* **2002**, *23*, 1623–1641.

[334] A. Ø. Madsen, S. Mason, S. Larsen, "A neutron diffraction study of xylitol: derivation of mean square internal vibrations for H atoms from a rigid-body description", *Acta Crystallogr. B* **2003**, *59*, 653–663.

[335] S. Aldridge, A. J. Downs, "Hydrides of the main-group metals: new variations on an old theme", *Chem. Rev.* **2001**, *101*, 3305–3366.

[336] J. R. Norton, J. Sowa, "Introduction: metal hydrides", *Chem. Rev.* **2016**, *116*, 8315–8317.

[337] S. Fukuzumi, T. Suenobu, "Hydrogen storage and evolution catalysed by metal hydride complexes", *Dalton Trans.* **2013**, *42*, 18–28.

[338] S. Niaz, T. Manzoor, A. H. Pandith, "Hydrogen storage: Materials, methods and perspectives", *Renew. Sustain. Energy Rev.* **2015**, *50*, 457–469.

[339] C. Mealli, T. B. Rauchfuss, "Models for the hydrogenases put the focus where it should be—Hydrogen", *Angew. Chem. Int. Ed.* **2007**, *46*, 8942–8944.

[340] D. Schilter, J. M. Camara, M. T. Huynh, S. Hammes-Schiffer, T. B. Rauchfuss, "Hydrogenase enzymes and their synthetic models: the role of metal hydrides", *Chem. Rev.* **2016**, *116*, 8693–8749.

[341] Peter Mueller, *Crystal structure refinement : a crystallographer's guide to SHELXL*, Reprinted, Oxford Univ. Press, Oxford, **2007**, p. 29.

[342] S. Pawlędzio, M. Malinska, M. Woińska, et al., "Relativistic Hirshfeld atom refinement of an organo-gold (I) compound", *IUCrJ* **2021**, *8*.

[343] P. Jadhav, G. Schatte, S. Labiuk, et al., "Cyclolinopeptide K butanol disolvate monohydrate", *Acta Crystallogr. E* **2011**, *67*, o2360–o2361.

[344] A. Belitzky, I. Weissbuch, Y Posner-Diskin, M. Lahav, I. Lubomirsky, "Design of Pyroelectric Mixed Crystals Having a Varying Degree of Polarity: The l-Asparagine· H2O/l-Aspartic Acid System", *Cryst. Growth Des.* **2015**, *15*, 2445–2451.

[345] E. Hupf, L. A. Malaspina, S. Holsten, et al., "Proximity Enforced Agostic Interactions Involving Closed-Shell Coinage Metal Ions", *Inorg. Chem.* **2019**, *58*, 16372–16378.

[346] R. J. Baker, A. J. Davies, C. Jones, M. Kloth, "Structural and spectroscopic studies of carbene and N-donor ligand complexes of Group 13 hydrides and halides", *Journal of organometallic chemistry* **2002**, *656*, 203–210.

[347] E. K. Wieduwilt, *Heavy meets light: A systematic study of the bond between heavy elements and hydrogen atoms*, Master thesis, University of Bremen, **2018**.

[348] V. Arion, J.-J. Brunet, D. Neibecker, "Crystal structure, Mössbauer spectra, and thermal behavior of H2Fe (CO) 2 [P (OPh) 3] 2", *Inorg. Chem.* **2001**, *40*, 2628–2630.

[349] N. N. Ho, R. Bau, S. A. Mason, "Neutron diffraction study of the highly distorted octahedral complex FeH2 (CO) 2 [P (OPh) 3] 2", *J. Organomet. Chem.* **2003**, *676*, 85–88.

[350] E. Arunan, G. R. Desiraju, R. A. Klein, et al., "Definition of the hydrogen bond (IUPAC Recommendations 2011)", *Pure Appl. Chem.* **2011**, *83*, 1637–1641.

[351] E. Arunan, G. R. Desiraju, R. A. Klein, et al., "Defining the hydrogen bond: An account (IUPAC Technical Report)", *Pure Appl. Chem.* **2011**, *83*, 1619–1636.

[352] L. Brammer, "Halogen bonding, chalcogen bonding, pnictogen bonding, tetrel bonding: origins, current status and discussion", *Faraday Discuss.* **2017**, *203*, 485–507.

[353] E. A. Meyer, R. K. Castellano, F. Diederich, "Interactions with aromatic rings in chemical and biological recognition", *Angew. Chem. Int. Ed.* **2003**, *42*, 1210–1250.

[354] L. M. Salonen, M. Ellermann, F. Diederich, "Aromatic rings in chemical and biological recognition: energetics and structures", *Angew. Chem. Int. Ed.* **2011**, *50*, 4808–4842.

[355] C. A. Hunter, J. K. Sanders, "The nature of. pi.-. pi. interactions", *J. Am. Chem. Soc.* **1990**, *112*, 5525–5534.

[356] D. A. Dougherty, "The cation- π interaction", *Accounts of chemical research* **2013**, *46*, 885–893.

[357] J. P. Gallivan, D. A. Dougherty, "Cation-π interactions in structural biology", *Proceedings of the National Academy of Sciences* **1999**, *96*, 9459–9464.

[358] A. Frontera, P. Gamez, M. Mascal, T. J. Mooibroek, J. Reedijk, "Putting anion–π interactions into perspective", *Angew. Chem. Int. Ed.* **2011**, *50*, 9564–9583.

[359] X. Lucas, A. Bauzá, A. Frontera, D. Quinonero, "A thorough anion–π interaction study in biomolecules: on the importance of cooperativity effects", *Chemical science* **2016**, *7*, 1038–1050.

[360] D. Nachtigallová, P. Hobza in *Complementary Bonding Analysis*, De Gruyter, **2021**, pp. 309–328.

[361] J. Černỳ, P. Hobza, "Non-covalent interactions in biomacromolecules", *Phys. Chem. Chem. Phys* **2007**, *9*, 5291–5303.

[362] M. K. Corpinot, D.-K. Bučar, "A practical guide to the design of molecular crystals", *Cryst. Growth Des.* **2018**, *19*, 1426–1453.

[363] P. Mignon, S. Loverix, J. Steyaert, P. Geerlings, "Influence of the π–π interaction on the hydrogen bonding capacity of stacked DNA/RNA bases", *Nucleic Acids Res.* **2005**, *33*, 1779–1789.

[364] C. N. Pace, J. M. Scholtz, G. R. Grimsley, "Forces stabilizing proteins", *FEBS Lett.* **2014**, *588*, 2177–2184.

[365] K. A. Dill, "Dominant forces in protein folding", *Biochemistry* **1990**, *29*, 7133–7155.

[366] K. A. Dill, J. L. MacCallum, "The Protein-Folding Problem, 50 Years On", *Science* **2012**, *338*, 1042–1046.

[367] S. Fiedler, J. Broecker, S. Keller, "Protein folding in membranes", *Cell. Mol. Life Sci.* **2010**, *67*, 1779–1798.

[368] E. Frieden, "Non-covalent interactions: Key to biological flexibility and specificity", *J. Chem. Educ.* **1975**, *52*, 754.

[369] S. K. Panigrahi, G. R. Desiraju, "Strong and weak hydrogen bonds in the protein–ligand interface", *Proteins: Struct. Funct. Bioinf.* **2007**, *67*, 128–141.

[370] A. J. Smith, X. Zhang, A. G. Leach, K. Houk, "Beyond picomolar affinities: quantitative aspects of noncovalent and covalent binding of drugs to proteins", *J. Med. Chem.* **2009**, *52*, 225–233.

[371] C. Bissantz, B. Kuhn, M. Stahl, "A medicinal chemist's guide to molecular interactions", *J. Med. Chem.* **2010**, *53*, 5061–5084.

[372] I. M. Nooren, J. M. Thornton, "Structural characterisation and functional significance of transient protein–protein interactions", *J. Mol. Biol.* **2003**, *325*, 991–1018.

[373] O. Keskin, N. Tuncbag, A. Gursoy, "Predicting protein–protein interactions from the molecular to the proteome level", *Chem Rev.* **2016**, *116*, 4884–4909.

[374] R. Raucci, E. Laine, A. Carbone, "Local interaction signal analysis predicts protein–protein binding affinity", *Structure* **2018**, *26*, 905–915.

[375] R. Laplaza, F. Peccati, R. A. Boto, et al., "NCIPLOT and the analysis of noncovalent interactions using the reduced density gradient", *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, e1497.

[376] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, M. Edelman, "Automated analysis of interatomic contacts in proteins.", *Bioinformatics (Oxford England)* **1999**, *15*, 327–332.

[377] I. K. McDonald, J. M. Thornton, "Satisfying hydrogen bonding potential in proteins", *J. Mol. Biol.* **1994**, *238*, 777–793.

[378] S. Contreras-Riquelme, J.-A. Garate, T. Perez-Acle, A. J. Martin, "RIP-MD: a tool to study residue interaction networks in protein molecular dynamics", *PeerJ* **2018**, *6*, e5998.

[379] G. J. Bartlett, A. Choudhary, R. T. Raines, D. N. Woolfson, "n$\rightarrow \pi^*$ interactions in proteins", *Nat. Chem. Biol.* **2010**, *6*, 615–620.

[380] G. J. Bartlett, R. W. Newberry, B. VanVeller, R. T. Raines, D. N. Woolfson, "Interplay of hydrogen bonds and n$\rightarrow \pi^*$ interactions in proteins", *J. Am. Chem. Soc.* **2013**, *135*, 18682–18688.

[381] M. M. Harding, M. W. Nowicki, M. D. Walkinshaw, "Metals in protein structures: a review of their principal features", *Crystallogr. Rev.* **2010**, *16*, 247–302.

[382] M. Gurusaran, M. Shankar, R. Nagarajan, J. R. Helliwell, K. Sekar, "Do we see what we should see? Describing non-covalent interactions in protein structures including precision", *IUCrJ* **2014**, *1*, 74–81.

[383] R. A. Klein, "Modified van der Waals atomic radii for hydrogen bonding based on electron density topology", *Chem. Phys. Lett.* **2006**, *425*, 128–133.

[384] A. L. Spek, "Single-crystal structure validation with the program PLATON", *J. Appl. Crystallogr.* **2003**, *36*, 7–13.

[385] P. R. Spackman, M. J. Turner, J. J. McKinnon, et al., "CrystalExplorer: a program for Hirshfeld surface analysis, visualization and quantitative analysis of molecular crystals", *J. Appl. Crystallogr.* **2021**, *54*.

[386] A. Bondi, "van der Waals volumes and radii", *J. Phys. Chem.* **1964**, *68*, 441–451.

[387] R. S. Rowland, R. Taylor, "Intermolecular nonbonded contact distances in organic crystal structures: Comparison with distances expected from van der Waals radii", *J. Phys. Chem.* **1996**, *100*, 7384–7391.

[388] I. Y. Chernyshov, I. V. Ananyev, E. A. Pidko, "Revisiting van der Waals Radii: From Comprehensive Structural Analysis to Knowledge-Based Classification of Interatomic Contacts", *ChemPhysChem* **2020**, *21*, 370.

[389] R. S. Paton, J. M. Goodman, "Hydrogen bonding and π-stacking: how reliable are force fields? A critical evaluation of force field descriptions of nonbonded interactions", *J. Chem. Inf. Model.* **2009**, *49*, 944–955.

[390] H. S. Muddana, M. K. Gilson, "Prediction of SAMPL3 host–guest binding affinities: evaluating the accuracy of generalized force-fields", *J. Comput. Aided Mol* **2012**, *26*, 517–525.

[391] P. Hobza, "Calculations on noncovalent interactions and databases of benchmark interaction energies", *Acc. Chem. Res.* **2012**, *45*, 663–672.

[392] N. D. Yilmazer, M. Korth, "Enhanced semiempirical QM methods for biomolecular interactions", *Comput. Struct. Biotechnol. J.* **2015**, *13*, 169–175.

[393] S. Grimme, "Density functional theory with London dispersion corrections", *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 211–228.

[394] S. Grimme, A. Hansen, J. G. Brandenburg, C. Bannwarth, "Dispersion-corrected mean-field electronic structure methods", *Chem. Rev.* **2016**, *116*, 5105–5154.

[395] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, "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.

[396] J. Rezác, K. E. Riley, P. Hobza, "S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures", *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.

[397] S. F. Boys, F. Bernardi, "The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors", *Mol. Phys.* **1970**, *19*, 553–566.

[398] M. J. Phipps, T. Fox, C. S. Tautermann, C.-K. Skylaris, "Energy decomposition analysis approaches and their evaluation on prototypical protein–drug interaction patterns", *Chem. Soc. Rev.* **2015**, *44*, 3177–3211.

[399] L. Zhao, M. von Hopffgarten, D. M. Andrada, G. Frenking, "Energy decomposition analysis", *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1345.

[400] E. Pastorczak, C. Corminboeuf, "Perspective: Found in translation: Quantum chemical tools for grasping non-covalent interactions", *J. Chem. Phys.* **2017**, *146*, 120901.

[401] B. Jeziorski, R. Moszynski, K. Szalewicz, "Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes", *Chem. Rev.* **1994**, *94*, 1887–1930.

[402] K. Szalewicz, "Symmetry-adapted perturbation theory of intermolecular forces", *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 254–272.

[403] H. L. Williams, C. F. Chabalowski, "Using Kohn- Sham orbitals in symmetry-adapted perturbation theory to investigate intermolecular interactions", *J. Phys. Chem. A* **2001**, *105*, 646–659.

[404] A. J. Misquitta, K. Szalewicz, "Intermolecular forces from asymptotically corrected density functional description of monomers", *Chem. Phys. Lett.* **2002**, *357*, 301–306.

[405] A. Heßelmann, G. Jansen, "The helium dimer potential from a combined density functional theory and symmetry-adapted perturbation theory approach using an exact exchange–correlation potential", *Phys. Chem. Chem. Phys.* **2003**, *5*, 5010–5014.

[406] S. E. Wheeler, K. Houk, "Through-space effects of substituents dominate molecular electrostatic potentials of substituted arenes", *J. Chem. Theory Comput.* **2009**, *5*, 2301–2312.

[407] P. Politzer, P. R. Laurence, K. Jayasuriya, "Molecular electrostatic potentials: an effective tool for the elucidation of biochemical phenomena.", *Environ. Health Perspect.* **1985**, *61*, 191–202.

[408] P. Politzer, J. S. Murray, Z. Peralta-Inga, "Molecular surface electrostatic potentials in relation to noncovalent interactions in biological systems", *Int. J. Quantum Chem.* **2001**, *85*, 676–684.

[409] S. E. Wheeler, J. W. Bloom, "Toward a more complete understanding of noncovalent interactions involving aromatic rings", *J. Phys. Chem. A* **2014**, *118*, 6133–6147.

[410] A. D. Becke, K. E. Edgecombe, "A simple measure of electron localization in atomic and molecular systems", *J. Chem. Phys.* **1990**, *92*, 5397–5403.

[411] F. R. Wagner, V. Bezugly, M. Kohout, Y. Grin, "Charge decomposition analysis of the electron localizability indicator: A bridge between the orbital and direct space representation of the chemical bond", *Chem. Eur. J* **2007**, *13*, 5724–5741.

[412] M. Alikhani, F Fuster, B Silvi, "What can tell the topological analysis of ELF on hydrogen bonding?", *Struct. Chem.* **2005**, *16*, 203–210.

[413] R. F. W. Bader, *Atoms in molecules : a quantum theory*, Clarendon Press, Oxford, **1990**.

[414] R. A. Boto, F. Peccati, R. Laplaza, et al., "NCIPLOT4: Fast, Robust, and Quantitative Analysis of Noncovalent Interactions", *J. Chem. Theory Comput.* **2020**, *16*, 4150–4158.

[415] C. Lefebvre, G. Rubez, H. Khartabil, et al., "Accurately extracting the signature of intermolecular interactions present in the NCI plot of the reduced density gradient versus electron density", *Phys. Chem. Chem. Phys.* **2017**, *19*, 17928–17936.

[416] C. F. Matta, R. J. Boyd in *The quantum theory of atoms in molecules: from solid state to DNA and drug design*, Wiley Online Library, **2007**, pp. 1–34.

[417] P. Popelier, "Characterization of a dihydrogen bond on the basis of the electron density", *J. Phys. Chem. A* **1998**, *102*, 1873–1878.

[418] R. F. W. Bader, "A bond path: a universal indicator of bonded interactions", *J. Phys. Chem. A* **1998**, *102*, 7314–7323.

[419] R. F. W. Bader, "Bond paths are not chemical bonds", *J. Phys. Chem. A* **2009**, *113*, 10391–10396.

[420] R. Laplaza, F. Peccati, D. Arias-Olivares, J. Contreras-García in *Complementary Bonding Analysis*, De Gruyter, **2021**, pp. 353–377.

[421] M. Blanco, A Martín Pendás, E Francisco, "Interacting quantum atoms: a correlated energy decomposition scheme based on the quantum theory of atoms in molecules", *J. Chem. Theory Comput.* **2005**, *1*, 1096–1109.

[422] E Francisco, A Martín Pendás, M. Blanco, "A molecular energy decomposition scheme for atoms in molecules", *J. Chem. Theory Comput.* **2006**, *2*, 90–102.

[423] P. Maxwell, Á. M. Pendás, P. L. Popelier, "Extension of the interacting quantum atoms (IQA) approach to B3LYP level density functional theory (DFT)", *Phys. Chem. Chem. Phys.* **2016**, *18*, 20986–21000.

[424] Á. M. Pendás, C. Gatti in *Complementary Bonding Analysis*, De Gruyter, **2021**, pp. 43–74.

[425] A. L. Wilson, P. L. Popelier, "Exponential relationships capturing atomistic short-range repulsion from the interacting quantum atoms (IQA) method", *J. Phys. Chem. A* **2016**, *120*, 9647–9659.

[426] J. Contreras-García, E. R. Johnson, S. Keinan, et al., "NCIPLOT: a program for plotting noncovalent interaction regions", *J. Chem. Theory Comput.* **2011**, *7*, 625–632.

[427] G. Saleh, C. Gatti, L. Lo Presti, J. Contreras-García, "Revealing Non-covalent Interactions in Molecular Crystals through Their Experimental Electron Densities", *Chem. Eur. J.* **2012**, *18*, 15523–15536.

[428] G. Saleh, C. Gatti, L. L. Presti, "Non-covalent interaction via the reduced density gradient: Independent atom model vs experimental multipolar electron densities", *Comput. Theor. Chem.* **2012**, *998*, 148–163.

[429] J. Contreras-García, R. A. Boto, F. Izquierdo-Ruiz, et al., "A benchmark for the non-covalent interaction (NCI) index or... is it really all in the geometry?", *Theor. Chem. Acc.* **2016**, *135*, 1–14.

[430] A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior", *Phys. Rev. A* **1988**, *38*, 3098.

[431] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange", *J. Chem. Phys* **1993**, *98*, 5648–6.

[432] A. Ghosh, A. Datta, J. Jana, et al., "Sequence context induced antimicrobial activity: insight into lipopolysaccharide permeabilization", *Mol. BioSyst.* **2014**, *10*, 1596–1612.

[433] M. F. Summers, T. L. South, B. Kim, D. R. Hare, "High-resolution structure of an HIV zinc fingerlike domain via a new NMR-based distance geometry approach", *Biochemistry* **1990**, *29*, 329–340.

[434] A. Langkilde, S. M. Kristensen, L. Lo Leggio, et al., "Short strong hydrogen bonds in proteins: a case study of rhamnogalacturonan acetylesterase", *Acta Crystallogr. D* **2008**, *64*, 851–863.

[435] M. J. Plevin, D. L. Bryce, J. Boisbouvier, "Direct detection of CH/π interactions in proteins", *Nat. Chem.* **2010**, *2*, 466.

[436] S. Vijay-Kumar, C. E. Bugg, W. J. Cook, "Structure of ubiquitin refined at 1.8 Å resolution", *J. Mol. Biol.* **1987**, *194*, 531–544.

[437] J. C. Ma, D. A. Dougherty, "The cation–π interaction", *Chem Rev.* **1997**, *97*, 1303–1324.

[438] A. E. Aleshin, B. Stoffer, L. M. Firsov, B. Svensson, R. B. Honzatko, "Crystallographic complexes of glucoamylase with maltooligosaccharide analogs: relationship of stereochemical distortions at the nonreducing end to the catalytic mechanism", *Biochemistry* **1996**, *35*, 8319–8328.

[439] B. Golinelli-Pimpaneau, "Structure of a pseudomerohedrally twinned monoclinic crystal form of a pyridoxal phosphate-dependent catalytic antibody", *Acta Crystallogr. D* **2005**, *61*, 472–476.

[440] R. W. Newberry, G. J. Bartlett, B. VanVeller, D. N. Woolfson, R. T. Raines, "Signatures of n$\rightarrow \pi^*$ interactions in proteins", *Protein Sci.* **2014**, *23*, 284–288.

[441] R. W. Newberry, R. T. Raines, "The n$\rightarrow \pi^*$ Interaction", *Acc. Chem. Res.* **2017**, *50*, 1838–1846.

[442] B. S. Avvaru, C. U. Kim, K. H. Sippel, et al., "A short, strong hydrogen bond in the active site of human carbonic anhydrase II", *Biochemistry* **2010**, *49*, 249–251.

[443] X. Dong, M. Kato-Murayama, T. Muramatsu, et al., "The crystal structure of leucyl/phenylalanyl-tRNA-protein transferase from Escherichia coli", *Protein Sci.* **2007**, *16*, 528–534.

[444] S. O. Dahms, I. Könnig, D. Roeser, et al., "Metal binding dictates conformation and function of the amyloid precursor protein (APP) E2 domain", *J. Mol. Biol.* **2012**, *416*, 438–452.

[445] T. Clark, J. Chandrasekhar, G. W. Spitznagel, P. V. R. Schleyer, "Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li-F", *J. Comput. Chem.* **1983**, *4*, 294–301.

[446] R. A. Kendall, T. H. Dunning, R. J. Harrison, "Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions", *J. Chem. Phys.* **1992**, *96*, 6796–6806.

[447] A. C. Wallace, R. A. Laskowski, J. M. Thornton, "LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions", *Protein Eng. Des. Sel.* **1995**, *8*, 127–134.

[448] R. A. Laskowski, M. B. Swindells, "LigPlot+: Multiple Ligand-Protein Interaction Diagrams for Drug Discovery", *J. Chem. Inf. Model.* **2011**, *51*, 2778–2786.

[449] K. Brejc, T. K. Sixma, P. A. Kitts, et al., "Structural basis for dual excitation and photoisomerization of the Aequorea victoria green fluorescent protein", *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 2306–2311.

[450] V. R. Kaila, R. Send, D. Sundholm, "Electrostatic spectral tuning mechanism of the green fluorescent protein", *Phys. Chem. Chem. Phys.* **2013**, *15*, 4491–4495.

[451] J. F. Amacher, P. R. Cushing, C. D. Bahl, T. Beck, D. R. Madden, "Stereochemical determinants of C-terminal specificity in PDZ peptide-binding domains: a novel contribution of the carboxylate-binding loop", *J. Biol. Chem.* **2013**, *288*, 5114–5126.

[452] Y. Zhao, D. G. Truhlar, "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 functionals", *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[453] M. Zimmer, "GFP: from jellyfish to the Nobel prize and beyond", *Chem. Soc. Rev.* **2009**, *38*, 2823–2832.

[454] R. Y. Tsien, "The green fluorescent protein", *Annu. Rev. Biochem.* **1998**, *67*, 509–544.

[455] J. Klein, H. Khartabil, J.-C. Boisson, et al., "New way for probing bond strength", *J. Phys. Chem. A* **2020**, *124*, 1850–1860.

[456] G. Münch, C. Dees, M. Hekman, D. Palm, "Multisite contacts involved in coupling of the β-adrenergic receptor with the stimulatory guanine-nucleotide-binding regulatory protein: Structural and functional studies by β-receptor-site-specific synthetic peptides", *Eur. J. Biochem.* **1991**, *198*, 357–364.

[457] D. Maiolo, A. Pizzi, A. Gori, et al., "Halogenation of the N-Terminus Tyrosine 10 Promotes Supramolecular Stabilization of the Amyloid-β Sequence 7–12", *ChemistryOpen* **2020**, *9*, 253.

[458] J. Klein, E. Pluot, J.-C. Boisson, E. Hénon, IGMPlot Revision 2.6.7, **2020**.

[459] M. J. Frisch, G. W. Trucks, H. B. Schlegel, et al., Gaussian 16, **2016**.

[460] I. W. Davis, A. Leaver-Fay, V. B. Chen, et al., "MolProbity: all-atom contacts and structure validation for proteins and nucleic acids", *Nucleic Acids Res.* **2007**, *35*, W375–W383.

[461] V. B. Chen, W. B. Arendall, J. J. Headd, et al., "MolProbity: all-atom structure validation for macromolecular crystallography", *Acta Crystallogr. D* **2010**, *66*, 12–21.

[462] C. J. Williams, J. J. Headd, N. W. Moriarty, et al., "MolProbity: More and better reference data for improved all-atom structure validation", *Protein Sci.* **2018**, *27*, 293–315.

[463] S. Grimme, S. Ehrlich, L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory", *J. Comput. Chem.* **2011**, *32*, 1456–1465.

[464] E. K. Wieduwilt, G. Macetti, R. Scatena, P. Macchi, A. Genoni, "Extending libraries of extremely localized molecular orbitals to metal organic frameworks: A preliminary investigation", *Crystals* **2021**, *11*, 207.

Résumé

Dans cette thèse, les orbitales moléculaires extrêmement localisées (ELMOs) ont été utilisées comme briques LEGO électroniques pour accomplir deux objectifs : (i) améliorer la précision des structures cristallines aux rayons X pour des petits et grands systèmes, et (ii) analyser les interactions non-covalentes dans les biomolécules. Les ELMOs sont des orbitales moléculaires strictement localisées sur des petits fragments. Pour cette raison, elles peuvent être calculées sur des petites molécules, stockées dans des bases de données, puis transférées sur des systèmes plus grands pour reconstruire leurs fonctions d'onde et leurs densités électroniques. Dans ce but, nous avons exploité les bibliothèques contenant les ELMOs de tous les fragments élémentaires (atomes, liaisons et groupes fonctionnels) des acides aminés. Dans les situations où une plus grande précision s'avère nécessaire, nous avons utilisé la technique de « embedding » QM/ELMO, dans laquelle la partie cruciale du système examiné est traitée à un niveau quantique supérieur tandis que le reste du système est décrit par des ELMOs gelées.

En ce qui concerne le premier des objectifs mentionnés ci-dessus, il est important de noter que les raffinements cristallographiques standards sont basés sur le « independent atom model » (IAM), qui estime la densité électronique à partir d'une somme de densités atomiques sphériques. Cependant, les longueurs des liaisons élément-hydrogène résultant des raffinements IAM sont systématiquement trop courtes. Une méthode qui résout ce problème est la technique « Hirshfeld atom refinement » (HAR), qui est basée sur le calcul direct de la densité électronique en utilisant des méthodes de mécanique quantique. Pour les petites molécules, il a été prouvé que la technique HAR donne des longueurs de liaison élément-hydrogène en très bon accord avec les valeurs de référence neutroniques. Toutefois, pour les grands systèmes, l'applicabilité de la méthode HAR traditionnelle est limitée, car les calculs entièrement QM sous-jacents deviennent trop coûteux. Pour pallier ce problème, dans le travail décrit dans cette thèse, les bibliothèques ELMO et les techniques QM/ELMO ont été couplées avec la méthode HAR dans le but de raffiner les structures de grands systèmes moléculaires, mais également pour obtenir des structures plus précises dans le cas de petites molécules. En outre, la nécessité ou non d'utiliser des méthodes post-HF pour HAR a été évaluée.

Le deuxième objectif de cette thèse concerne l'analyse des interactions non covalentes, qui pose des problèmes similaires à ceux rencontrés en cristallographie aux rayons X. En effet, les techniques « non-covalent interaction » (NCI) et « independent gradient model » (IGM), qui sont couramment appliquées dans l'analyse des interactions non covalentes, dépendent du calcul de la densité électronique. Ainsi, pour analyser les interactions non covalentes dans des grands systèmes, les deux techniques ont recours aux mêmes densités électroniques promoléculaires utilisées dans le « independent atom model ». Cependant, dans le cas des analyses NCI et IGM, ces densités fournissent également des résultats biaisés. Pour surmonter cet inconvénient, nous avons couplé les deux méthodes avec les bibliothèques ELMO, donnant naissance aux techniques NCI-ELMO et IGM-ELMO, qui ont ensuite été appliquées pour identifier, classer et quantifier approximativement les interactions non covalentes dans des polypeptides et des protéines.

Abstract

In this thesis, extremely localized molecular orbitals (ELMOs) were used as electronic LEGO building blocks to accomplish mainly two goals: (i) obtaining more accurate X-ray crystal structures for small and large systems, and (ii) analyzing non-covalent interactions in biomolecules. In fact, ELMOs are molecular orbitals that are strictly localized on small molecular fragments. Due to this strict localization, they may be computed on small molecules, stored in databases and then transferred to larger systems to reconstruct their wavefunctions and electron densities. To this end, we exploited the ELMO libraries, which contain the ELMOs for all the elementary fragments (atoms, bonds and functional groups) of the twenty natural amino acids. In situations where a higher accuracy was needed, we used the QM/ELMO embedding technique, in which the crucial part of the system under exam is treated at a higher quantum mechanical level, while the rest is described using frozen ELMOs.

Concerning the first of the goals mentioned above, it is important to note that standard crystallographic refinements are based on the so-called independent atom model (IAM), which approximates the electron density as a sum of spherically averaged atomic densities. However, the element-hydrogen bond lengths resulting from IAM refinements are systematically too short. A method that solves this problem is the Hirshfeld atom refinement (HAR), a technique based on directly computing the electron density for the molecule under exam using quantum mechanical calculations. For small molecules, HAR has been proven to give element-hydrogen bond lengths that are in very good agreement with neutron reference values. However, for large systems, the applicability of the traditional HAR method is limited because the underlying fully QM calculations become computationally too expensive. Therefore, in the work presented in this thesis, the ELMO libraries and the QM/ELMO techniques have been coupled with the HAR method to refine large systems and also to obtain more accurate structures of small molecules. Furthermore, the necessity of using post-HF methods for HAR has been also evaluated.

Concerning the second goal addressed in this dissertation, a similar problem as the one met in X-ray crystallography also arises in the analysis of non-covalent interactions. In fact, also the non-covalent interaction (NCI) and independent gradient model (IGM) techniques, which are commonly applied in analyses of non-covalent interactions, crucially depend on the computation of the electron density. Therefore, to analyze non-covalent interactions in large systems, both techniques had to resort to promolecular electron densities, which are the same densities used in the IAM. However, also in the cases of NCI and IGM analyses, these densities provide biased results. To overcome this drawback, we have coupled both methods with the ELMO libraries, giving rise to the NCI-ELMO and IGM-ELMO techniques, which were then applied to identify, classify and approximately quantify non-covalent interactions in polypeptides and proteins.