Institute of Physical Chemistry
Polish Academy of Sciences
Kasprzaka 44/52
01-224 Warsaw, Poland

PhD Thesis

Density Functional Theory and Information Theory Based Indices as Tools to Investigate the Reactivity of Chemical Systems and Their Applications

Meressa Abrha Welearegay

Supervisor: Prof. dr. hab. Andrzej Holas

Subsidiary supervisor: Dr. Robert Balawender

This dissertation was prepared within the International PhD in Chemistry Studies at the Institute of Physical Chemistry of the Polish Academy of Sciences in Warsaw.

Project operated within the Foundation for Polish Science International PhD Projects Programme co-financed by the European Regional Development Fund, Operational Program Innovative Economy 2009-2013.

INNOWACYJNA GOSPODARKA
NARODOWA STRATEGIA SPÓJNOŚCI

FNP
Fundacja na rzecz Nauki Polskiej

UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ
ROZWOJU REGIONALNEGO

Warsaw, May 2014

Acknowledgments

First and foremost, I would like to thank my supervisors Prof. Andrzej Holas and Dr. Robert Balawender for their tireless effort and motivation to come this thesis a reality. I greatly appreciate and thankful for their willingness to help me in scientific problems and other social matters.

I would like to acknowledge the Institute of Physical Chemistry Polish Academy of Science and its community for making friendly atmosphere to accomplish this work. Thanks to all colleagues and people of the quantum chemistry staff whom I have had the delight to interact.

I am grateful to Prof. Paul Ayers for offering the opportunity to visit his group at McMaster University, Hamilton, Canada. It was a great experience working with him and his helpful discussions on scientific problems greatly influenced my way of research direction. Thanks to all of his staff and the department of Chemistry and Chemical Biology.

I am grateful to Prof. Paul Geerlings for his amazing guidance and helpful discussions during my stay in his research team at the Vrije Universiteit Brussel, Brussels, Belgium. I also appreciate the advice and help from Prof. De Proft. I want to extend my appreciation to their sociable and helpful staff and the department of General Chemistry (ALGC).

Last but not least, I would like to thank to my family, relatives and friends for their great courage and motivation.

“አምላኬ ሆይ ብርታቱን ስለሰጠሀኝ አመሰግናሃለሁ።”

Acknowledgments

I would like to thank the International PhD Projects Programme of the Foundation for Polish Science, co-financed from European Regional Development Fund within Innovative Economy Operational Programme “Grants for Innovation” MPD/2009/1/styp8/ for financial support.

I would also like to thank

the (ICM) Interdisciplinary Center for Mathematical and Computational Modeling, Poland; the compute Canada SHARCNET (Shared Hierarchical Academic Research Computer Network; the VUB Computer Center (Hydra), Belgium

for computational grants.

Abstract

A general theme of this dissertation is showing usefulness of the density functional theory (DFT) and the information theory (IT) based indices as a source and carrier of the information about molecular structure and reactivity.

One of the main targets undertaken in this work is opening the gate to the exploration of the chemical space through alchemical derivatives. This is done by developing the methodological framework of alchemical derivatives and their chemical reactivity based indices. The crucial conditions for the qualitative and quantitative accuracies of the alchemical predictions are recognized. As illustrative transmutations showing the potential of the method in exploring chemical space, some examples of increasing complexity are presented, starting with the deprotonation, continuing with the transmutation of the nitrogen molecule, and ending with the substitution of isoelectronic (B,N) units for (C,C) units and N units for C-H units in carbocyclic systems. The overall trends observed for the alchemical deprotonation energy prove the usefulness of the alchemical indices as the probe in the chemical reactivity investigations. The results of calculations for the BN derivatives of benzene and pyrene show that this method has great potential for efficient and accurate scanning of chemical space. The tentative results for the carcinogenic activity of the polycyclic aromatic hydrocarbons (PAHs) reinforce this opinion.

The information theory based methodological framework necessary for extracting useful information for atomic and molecular systems is deeply examined. The concepts of the transferability and additivity of atoms or functional groups is used as a test in extensive and detailed analyses of the electronic density and the shape function as a functional argument for the IT based measures. It is also shown that in the shape representation, the observed trends for the IT measures and complexities are in contradiction with chemical intuition.

It is important to point out that linear correlations are obtained between the kinetic energy and the Fisher information and Onicescu information measures as well as between the atomization energy and the atomization entropy. The analysis of the IT based measures of information planes shows that the Shannon-Fisher information plane provides “richer” information about the pattern, organization, similarity of molecules than the Shannon-Onicescu and Fisher-Onicescu planes. The final conclusion for this part is that, the IT measures can be used in the chemical reactivity investigation as the source of the information about the pattern, organization, similarity of molecules, while not as direct indicators of their reactivity.

Finally, a support vector machine based models of classification and regression are developed for the carcinogenic effect of PAHs. The accuracy of 93% of correct classification is achieved using selected structural and molecular descriptors. The correlation coefficient for the predicted versus experimental index of carcinogenicity is 0.9475.

The used set of molecule is large enough (in total around 1000 molecules) and diverse to improve the previous understating of subject undertaken in this thesis and to generalize obtained conclusions.

Streszczenie

Ogólnym tematem tej rozprawy jest pokazanie użyteczności wskaźników opartych na teorii funkcjonałów gęstości (DFT) i teorii informacji (IT) jako źródła i nośnika informacji o strukturze i reaktywności cząsteczek.

Jednym z głównych celów podjętych w tej pracy jest zainicjowanie wykorzystania pochodnych alchemicznych w badaniu „przestrzeni związków chemicznych". W tym celu zdefiniowano metodologię pochodnych alchemicznych oraz indeksów na nich bazujących. Określono wpływ funkcji bazy na jakość otrzymywanych wyników, zarówno ilościowych, jak i jakościowych. Jako przykładowe transmutacje pokazujące możliwości tej metody, szereg przykładów o wzrastającej złożoności został przebadany: reakcja deprotonacji, transmutacji cząsteczki azotu, podstawienia pary węgli przez atom boru i azotu oraz zamiana grupy C-H na atom azotu. Obserwowane trendy dla energii deprotonacji potwierdzają użyteczność wskaźników alchemicznych w badaniach reaktywności chemicznej. Wyniki obliczeń pochodnych BN benzenu i pirenu pokazują duży potencjał tej metody w efektywnym i precyzyjnym przeszukiwaniu przestrzeni związków chemicznych. Wstępne wyniki badania aktywności rakotwórczej policyklicznych węglowodorów aromatycznych (PAH) wzmacniają tę opinię.

Szczegółowo przebadano metodologię opisywania własności atomowych i cząsteczkowych przy użyciu teorii informacji. Miary teorii informacji, będące funkcjonałami gęstości elektronowej lub funkcji kształtu, zostały szeroko i szczegółowo przebadane z wykorzystaniem pojęć przenoszalności i addytywności atomów i grup funkcyjnych. Pokazano, że użycie funkcji kształtu, jako argumentu funkcjonalnego prowadzi do wyników sprzecznych z „intuicją" chemiczną.

Warto podkreślić otrzymanie linowych korelacji pomiędzy energią kinetyczną a informacja Fishera i informacja Onicescu oraz pomiędzy energia atomizacji a entropia atomizacji. Analiza płaszczyzn informacyjnych pokazała, że najbogatsza informacja o wzorcach, organizacji, podobieństwach cząsteczek jest zawarta w płaszczyźnie informacji Shanona-Fishera niż w płaszczyznach Shannona-Onicescu czy Fisher-Onicescu. Stwierdzono że miary IT mogą być użyte w badaniach reaktywności chemicznej jako źródło informacji o strukturze , organizacji, podobieństwa cząsteczek, lecz nie jako bezpośrednie wskaźniki ich reaktywności.

W ostatniej części zaproponowano modele klasyfikacji rakotwórczej aktywności PAH przy wykorzystaniu metody maszyn wektorów nośnych (support vector machines). Bazując na wskaźnikach związanych z budową cząsteczek i jej własnościami otrzymano 93 % dokładność klasyfikacji. Współczynnik korelacji pomiędzy przewidywanymi a eksperymentalnymi wskaźnikami rakotwórczości wyniósł o 0,9475.

Wykorzystany w pracy zestaw cząsteczek był na tyle duży i różnorodny, że pozwolił na lepsze zrozumienie badanych zależności oraz na uogólniające wnioski.

Table of Contents

List of Tables

List of Figures

Abbreviations

1-DM	First-order density matrix
2-DM	Second-order density matrix
a.u.	Atomic unit
ADA	Adaptive boosting
AO	Atomic orbital
CA	Carcinogenic activity
CC	Coupled-cluster
cc-pVnZ	Correlation-consistent polarized valence number of zeta
cc-pCVnZ	As cc-pVnZ with addition of tight functions
CCSD(T)	Coupled cluster for singles/doubles/selected triples coupling terms
CI	Configuration interaction
CP	Carcinogenic potency
CS	Chemical space
CV	Cross-validation
DFA	Density functional approximation
DFRT	Density functional reactivity theory
DFT	Density function theory
DM	Density matrix
EA	Electron affinity
FCI	Full configuration interaction
FI	Fisher information
FN	False negative
FP	False positive
FPR	False positive rate
FS	Fisher-Shannon complexity
GAMESS	The general atomic and molecular electronic structure system
GGA	General gradient approximation
GS	Ground state
GTO	Gaussian-Type Orbitals
HF	Hartree-Fock
HK	Hohenberg-Kohn
HOMO	Highest occupied molecular orbital

HSAB	Pearson's hard-soft and acid-base
IE	Vertical ionization energy
IT	Information theory
KS	Kohn-Sham
LCAO	Linear combination of atomic orbitals
LDA	Local density approximation
LMC	López-Ruiz–Mancini–Cablet
LUMO	Lowest unoccupied molecular orbital
MAE	The mean absolute error
MAPE	The mean absolute percentage error
MO	Molecular orbital
N-DM	N -electron density matrix
OI	The Onicescu information
PAHs	Polycyclic aromatic hydrocarbons
Ref	Reference
SE	Shannon entropy
SID	Support information disc (see After Appendix C)
SVM	Support vector machine
TN	True negative
TP	True positive
TPR	True positive rate

Chapter I. Introduction

The analysis and investigation of atoms and molecules (as building blocks of matter) have a vital role in our daily basis. This work focuses on the basic ideas, concepts and methodology along with their interesting results for various chemical systems that readers can easily understand and know the current challenges and efforts that had been done (or those in due process) that can bring new insight in the scientific community.

The general aim of this dissertation is to verify the usefulness of the density functional theory (DFT) and the information theory (IT) based indices as tools to investigate the properties of chemical systems. In more elaborated way, apart from Chapter II. *Theoretical Background* where overall computational details are described and the summary (conclusion) part, the work is divided into four parts:

Chapter III. *Chemical Space and Reactivity Indices* is devoted to develop the methodological framework of alchemical derivatives and their chemical reactivity indices within the chemical space.

The purpose of Chapter IV. *Examples and Applications of Alchemical Derivatives* is to demonstrate and test the alchemical derivative applications within the chemical space. The exploration of stable molecules from the chemical space is mainly investigated. So, the first task is to find the best computational method (to study the basis set and functional dependence) for computing these derivatives. Next is to examine explicitly the transmutation energy and its components allowing to distinguish among different stable transmutation products (transmutants or isomers). Moreover, the effect of substituents on the deprotonation energy and its component contributions will be examined.

Chapter V. *Information and Complexity Measures in the Molecular Reactivity Studies* is aimed to explore the IT measures and complexities of varieties of molecules. Here, the goal is at first to explicitly examine the spinor and spinless densities and their corresponding shape functional arguments so as to decide which functional argument and to which chemical system should be recommended so that more information could be extracted for a given atom or molecule. It is also important to examine the behavior of individual information measures (such as the Shannon entropy, Fisher information, and Onicescu information (disequilibrium)), and the complexity measures. Furthermore, investigating the relationship among the IT reactivity measures and DFT chemical reactivity indices is the other theme is discussed. Another purpose is to establish a mathematical formalism for the relations among different types of densities and shape functionals. The analysis and comparison of atomic forms of IT measures and /or energies (e.g. atomization energy) with their molecular forms will be tested.

The goal of Chapter VI. *Analysis of Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons* is to investigate the carcinogenic activity (CA) of polycyclic aromatic hydrocarbons (PAHs) in two ways. First, the carcinogenicity database will be constructed. Peripheral and E-Dragon descriptors (using E-Dragon software) will be used to develop the support vector machine (SVM) based model of carcinogenicity. The E-Dragon software is an application for the calculation of molecular descriptors. These descriptors can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and high throughput screening of molecule databases.[1,2] The SVM is a supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. Three SVM classification models will be proposed and tested. Secondly, the alchemical deprotonation energy of one or two proton is used in the qualitative investigation of the CA of PAHs.

Chapter II. Theoretical Background

In quantum chemistry, quantum mechanics is applied to chemical problems. In most cases, one tries to solve the (non-relativistic) time-independent Schrödinger equation, leading to (quantized) energies of the system and the associated wave functions. However, only the simplest chemical systems can be described exactly in a quantum mechanical way. Indeed, the large number of particles that have to be treated in a chemical system is certainly one of the greatest obstacles to this way. In practice, approximations are introduced, resulting in the construction of methods such as Hartree-Fock, post-Hartree-Fock and Density Functional Theory (DFT). The latter method uses, instead of the wave function, the electron density as the basic source of information on an atomic or a molecular system. Its low computational requirements, compared to more traditional *ab initio* wavefunction methods, allow for application to large systems, which has made it the computational workhorse suitable.

II.A Quantum chemistry

II.A.1 Schrödinger equation and variational principle

In quantum chemistry (in most cases), one tries to find (approximate) solutions to the time-independent, non-relativistic Schrödinger equation:

$$\hat{H}\Psi_I\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N,\mathbf{R}_1,\mathbf{R}_2,\cdots,\mathbf{R}_P\right)=E_I\Psi_I\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N,\mathbf{R}_1,\mathbf{R}_2,\cdots,\mathbf{R}_P\right).\quad(2\text{-}1)$$

E_I is the energy of the I^{th} state described by wavefunction Ψ_I. Here $\mathbf{x}_i=\left(\mathbf{r}_i,\kappa_i\right)$ consists of spatial variable $\mathbf{r}_i$ and spin variable $\kappa_i\in\{\uparrow,\downarrow\}\equiv\{1/2,-1/2\}$ of the electron, while $\mathbf{R}_A$ is the coordinate of the nucleus. $\hat{H}$ is the molecular Hamiltonian of a system containing P nuclei (indices A and B) and N electrons (indices i and j) in the absence of magnetic or electric fields (using atomic units, $|e|=\hbar=m_{\text{e}}=1$ a.u.)

$$\hat{H}=\hat{T}_{\text{e}}+\hat{T}_{\text{n}}+\hat{V}_{\text{ne}}+\hat{V}_{\text{ee}}+\hat{V}_{\text{nn}}.\qquad(2\text{-}2)$$

Its terms are

– the kinetic energy operator of the electrons

$$\hat{T}_{\text{e}}=-\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2\qquad(2\text{-}3)$$

–the kinetic energy of the nuclei

$$\hat{T}_{\mathrm{n}} = -\frac{1}{2}\sum_{A=1}^{P}\frac{\nabla_A^2}{M_A} \tag{2-4}$$

– the coulomb attraction between the electrons and nuclei

$$\hat{V}_{\mathrm{ne}} = -\sum_{i=1}^{N}\sum_{A=1}^{P}\frac{Z_{\mathrm{A}}}{r_{iA}} \tag{2-5}$$

– the repulsion between electrons

$$\hat{V}_{\mathrm{ee}} = \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} \tag{2-6}$$

– the repulsion between nuclei

$$\hat{V}_{\mathrm{nn}} = \sum_{A=1}^{P}\sum_{B>A}^{P}\frac{Z_A Z_B}{R_{AB}}\,. \tag{2-7}$$

Here, $r_{ij} = \left|\mathbf{r}_i - \mathbf{r}_j\right|$, $r_{iA} = \left|\mathbf{r}_i - \mathbf{R}_A\right|$ and $R_{AB} = \left|\mathbf{R}_A - \mathbf{R}_B\right|$ are the distances, ∇_i^2 and ∇_A^2 are the Laplacian operators for the i^{th} electron and A^{th} nucleus having mass M_A and atomic number (charge) Z_{A}.

According to the Born-Oppenheimer approximation, the large difference in masses of electrons and nuclei allows for the separation of electronic and nuclear degrees of freedom (electron motion in the field of fixed nuclei). Hence, the electronic Hamiltonian extracted from Eq. (2-2) is

$$\hat{H}_{\mathrm{e}} = \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{ne}} + \hat{V}_{\mathrm{ee}}\,, \tag{2-8}$$

and the electronic Schrödinger equation is

$$\hat{H}_{\mathrm{e}}\Psi_{\mathrm{e}}\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\right) = E_{\mathrm{e}}\Psi_{\mathrm{e}}\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\right), \tag{2-9}$$

with $\Psi_{\mathrm{e}}\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\right)$being an antisymmetric electronic wavefunction with N spin-space coordinates $\mathbf{x}_i$. Note that nuclear coordinates enter $\hat{H}_{\mathrm{e}}$ parametrically through $\hat{V}_{\mathrm{ne}}$, resulting in the parametric dependence on nuclear coordinates in $\Psi_{\mathrm{e}} = \Psi_{\mathrm{e}}\left(\left\{\mathbf{x}_i\right\};\left\{\mathbf{R}_A\right\}\right)$ as well as in $E_{\mathrm{e}} = E_{\mathrm{e}}\left(\left\{\mathbf{R}_A\right\}\right)$[3]. From now Ψ will denote the electronic wavefunction (stated unless otherwise). The total energy of the system

$$W\left(\left\{\mathbf{R}_A\right\}\right)=E_{\mathrm{e}}\left(\left\{\mathbf{R}_A\right\}\right)+V_{\mathrm{nn}}\left(\left\{\mathbf{R}_A\right\}\right) \tag{2-10}$$

is the potential energy for the determination of the equilibrium configuration and vibrations of the molecule.

Unfortunately, except the one-electron systems, no strategy exists to solve the Schrödinger equation, Eq.(2-9) *exactly* for atomic or molecular systems. Therefore one tries to approach systematically to the wavefunction Ψ_{gs} of the system in the ground state (GS), which corresponds to the state with the lowest energy (i.e. the GS energy) E_0, through the *variational principle*[4,5]. Variational principle states that the energy of a trial wavefunction Ψ_{trial} is always greater than or equal to the true GS energy:

$$\begin{aligned}\left\langle\hat{H}\right\rangle_{\Psi_{\mathrm{trial}}} &\equiv \frac{\left\langle\Psi_{\mathrm{trial}}\middle|\hat{H}\middle|\Psi_{\mathrm{trial}}\right\rangle}{\left\langle\Psi_{\mathrm{trial}}\middle|\Psi_{\mathrm{trial}}\right\rangle}=\frac{\int\Psi_{\mathrm{trial}}{}^{*}\hat{H}\Psi_{\mathrm{trial}}d\tau}{\int\Psi^{*}\Psi d\tau} \\ &=E\left[\Psi_{\mathrm{trial}}\right]\geq E\left[\Psi_{\mathrm{gs}}\right]=\frac{\left\langle\Psi_0\middle|\hat{H}\middle|\Psi_0\right\rangle}{\left\langle\Psi_0\middle|\Psi_0\right\rangle}=E_0\end{aligned} . \tag{2-11}$$

Thus, $E\left[\Psi_{\mathrm{trial}}\right]$ is the upper bound to the GS energy E_0. The equality holds when the trial wavefunction, Ψ_{trial} is the same as the exact GS wave function, Ψ_0. Let us take the trial Ψ (subscript is omitted) expanded in a basis of K_{b} linearly independent functions, $\left\{\Phi_i\right\}$

$$\Psi=\sum_{i=1}^{K_{\mathrm{b}}}C_i\left\{\Phi_i\right\}. \tag{2-12}$$

Inserting it into Eq.(2-11) leads to a linear variational method (by minimization of):

$$E\left[\Psi\right]=\frac{\sum_{i,j}C_i^{*}C_j\left\langle\Phi_i\middle|\hat{H}\middle|\Phi_j\right\rangle}{\sum_{i,j}C_i^{*}C_j\left\langle\Phi_i\middle|\Phi_j\right\rangle}=\frac{\sum_{i,j}C_i^{*}C_jH_{ij}}{\sum_{i,j}C_i^{*}C_jS_{ij}} . \tag{2-13}$$

The $H_{ij}=\left\langle\Phi_i\middle|\hat{H}\middle|\Phi_j\right\rangle$ and $S_{ij}=\left\langle\Phi_i\middle|\Phi_j\right\rangle$ are the Hamiltonian and overlap matrix elements, respectively. $\left\{C_i\right\}$ are varied until $E\left[\Psi\right]$ get minimized. The necessary conditions, $\frac{\partial E\left[\Psi\right]}{\partial C_k^{*}}=0$ for all k, result in the secular equation

$$\sum_{j=1}^{K_\mathrm{b}}\left(H_{kj}-ES_{kj}\right)C_j=0, \qquad k=1,2,\cdots,K_\mathrm{b}. \tag{2-14}$$

This is $K_\mathrm{b}\times K_\mathrm{b}$ matrix secular equation. Hence, the solution of Eq.(2-14) is the variational estimate of the GS energy (and the energies of the first $K_\mathrm{b}-1$ excited states). Once the expansion coefficients, $\left\{C_i\right\}$, are calculated, Ψ is determined as the approximate GS function and thereby all properties can be obtained by evaluating expectation values of appropriate operators with this wave function, $\left\langle\Psi\middle|\hat{O}\middle|\Psi\right\rangle$.

Now we will try to find a subset of acceptable trial wave functions for which all calculations can be carried out easily and efficiently. One possibility is to build trial functions from products of one-electron functions, the molecular orbitals (MOs). Excellent reviews of *ab initio* wavefunction methods can, among others, be found in the following textbooks[4-6]:

II.A.2 Molecular orbital and basis set

Most calculational methods of quantum chemistry represent the N-electron wavefunction as a linear combination of Slater determinants, constructed of one-electron functions — molecular spin orbitals. They are written in terms of atomic spin orbitals or similar functions. An atomic orbital is labeled by quantum numbers: the principal (*n*), orbital (azimuthal) angular momentum (*l*), magnetic (l_z) and spin (s_z) quantum numbers[5,7]. The atomic spin orbital is a product of spatial orbital and an appropriate spin function: $\alpha\left(\kappa\right)$ when $s_z=1/2$ or $\beta\left(\kappa\right)$ when $s_z=-1/2$.

A molecular orbital is approximated as a linear combination (LC) of atomic orbitals (AOs) — (quantum superposition of atomic orbitals) LCAOs. For a given basis set of atomic orbitals $\left\{\chi_\mu\right\}$, the i^{th} molecular orbital ψ_i is

$$\psi_i=\sum_{\mu=1}^{K}c_{\mu,i}\chi_\mu\,;\ \mu=1,2,\cdots,K, \tag{2-15}$$

with $c_{\mu,i}$ the coefficients of expansion, and *K* the total number basis functions (atomic orbital functions).

The choice of appropriate basis set has a paramount importance in quantum chemical calculations, because the quality of all the results obtained will ultimately depend on the quality of this basis set.

For choosing the basis functions, two guidelines are necessary. [4] As a first point, they should reproduce the physics of the problem that ensures rapid convergence as the number of basis function increases. In particular, these functions should vanish at large distance from the nucleus. Secondly, the practical way of selecting elegant and good functions should allow for fast calculation of all integrals of interest (to save computational time).

The two commonly used types of atomic orbital functions that are applied in the LCAO method, Eq.(2-15), are the Slater-Type Orbitals (STO) [8] and Gaussian-Type Orbitals (GTO)[9] . STOs have a form

$$\chi^{\mathrm{STO}}_{\zeta,l,m}\left(r,\theta,\varphi\right)=\mathcal{N}\ r^{l}\exp\left(-\zeta r\right)Y_{lm}\left(\theta,\varphi\right), \tag{2-16}$$

which is the product of radial part $r^{l}\exp\left(-\zeta r\right)$, the angular part (the spherical harmonic functions $Y_{lm}\left(\theta,\varphi\right)$) and the normalization constant $\mathcal{N}$. The *l* and *m* are the atomic quantum numbers while *r* and ζ are the radial distance and the orbital exponent, respectively. The orbital exponent governs the size of the orbital, it is chosen for each (*n*, *l*) separately, where *n* is the principal quantum number. The GTOs, in terms of the Cartesian and polar coordinates have the form[4]

$$\chi^{\mathrm{GTO}}_{\zeta,l_x,l_y,l_z}\left(x,y,z\right)=\mathcal{N}\ x^{l_x}y^{l_y}z^{l_z}\exp\left(-\zeta r^{2}\right), \tag{2-17}$$

$$\chi^{\mathrm{GTO}}_{\zeta,l,m}\left(r,\theta,\varphi\right)=\mathcal{N}\ Y_{lm}\left(\theta,\varphi\right)r^{l}\exp\left(-\zeta r^{2}\right), \tag{2-18}$$

with $l_x+l_y+l_z=l$ that determines the type of orbital, such as $l=0,1,2$ represents the s, p and d orbitals, respectively.

In general, STOs are good and more convenient when high accuracy is needed for atomic and diatomic systems (*ab initio* methods), or where all three- and four-centre integrals are neglected (semi-empirical methods) as well as in cases where Coulomb energy is calculated by fitting the density into a set of auxiliary functions instead of computing the exact exchange energy (in DFT). However, use of STOs is time-consuming for computing the two-electron molecular integrals and difficult to be differentiated analytically. GTOs on the other hand, are more amenable to computation (fast molecular integrals) and can be differentiated trivially any number of times. Moreover, STO and GTO basis functions have different behavior near the nucleus: GTOs have no cusp at the origin while STOs have. The decay (for large r) of GTO differs from STO due to the squared distance in its exponent.

Apart from the choice of STO or GTO type of functions, the most important factor is the number of basis functions used in the basis set. Based on this, the commonly used basis sets can be classified into the following types:

– minimum basis set

The smallest number of basis functions per atom to describe the occupied atomic orbitals of that atom is known as the minimum basis set (relatively inexpensive in calculating quite large molecules). STO-nG (n being the number of contracted Gaussian functions, usually have values from 1 to 6) is an example of a basis function entering the minimum basis set. Minimum basis set gives accurate qualitative information such as chemical bonding. Usually, STO-3G functions are sufficient for minimum basis calculations (the computational cost increases with n).

– split valence (Pople) basis sets

The core (inner-shell) electrons of an atom are less affected by the chemical environment than the valence-shell electrons. Split valence basis set [4,10] also called Pople basis sets, can be described in such a way that the core shells are treated with a minimal basis set while the valence shells are treated with a larger basis set. In other words, each orbital of a core-shell is represented by one function, while two or more functions are used to represent each valence shell AO. The acronym for the split valence basis sets have a general form *q*-*rs*G or *q*-*rst*G, where *q* represents the number of primitives used in the contracted core functions, while *rs* or *rst* (after the hyphen) indicate the numbers of primitives used in the contracted valence functions: *rs* indicate presence of a two-function basis; *rst* — a three-function basis for a valence electron, the so called valence-double zeta and valence-triple zeta basis. Addition of "+" before G indicates diffusion function while addition of "*" after G indicates presence of polarization functions for heavy atoms. If, instead, "++" is added before G, it indicates diffuse functions added to both heavy and light atoms (hydrogen). Likewise, addition of "**" after G shows polarization functions also added to light atoms. Note that the s- and p-functions in the valence have the same exponent and this increase efficiency but decrease flexibility of basis sets.

Polarization functions are included to improve the flexibility of the basis set, especially for representing the electron density in bonding regions of a molecule. On the other hand, diffuse functions have small orbital exponent that makes the electron distribution very broad. Anions are well treated using these diffuse basis functions.

–correlation-consistent basis sets

The most frequently used basis sets are the Dunning-Huzinaga type basis sets or formally known as correlation consistent polarized valence Zeta, abbreviated as cc-pVnZ and aug-cc-pVnZ , where n = D (double), T (triple), Q (quadruple), 5, 6. Moreover, to describe the core-core or tight functions along with the core–valence correlation effects in atoms and molecules, the acronym "cc" is added. Presence of "aug" mean "augmented" which is equivalent to the "+" in Pople type basis that indicate the addition of diffusion functions to the basis set. When the complete basis set limit is investigated, the cc-pCV*n*Z basis sets allow for a systematic convergence to the electron correlation energy. The effect of adding the core electrons on the calculation of molecular properties tends to be dominated by core–valence correlation effects. This makes the core–valence correlation energy converging more slowly than the core-core correlation energy[11] .Unlike the Pople style split-valence**,** the correlation-consistent type basis sets avoid the restriction of using equal exponents for the s- and p-type functions in the valence shell. These exponents and contraction coefficients are variationally optimized for both Hartree-Fock and electron correlation calculations.

In general, even though it is difficult to construct one universal molecular basis set criteria that are applicable under all circumstances, there are three requirements that proper basis set should meet[12]

- the basis should be designed in such a way that it allows for a systematic saturation of increasingly higher angular momentum functions.
- the basis should yield a fast convergence of the self-consistent cycles.
- the basis should provide for an easy manipulation and efficient implementation of all the basic molecular integrals required. Also, the basis should not cause numerical instability problems.

II.B *Ab initio* molecular orbital methods

*Ab initio (*Latin phrase to mean *'from first principles'* or *'from beginning'*) molecular orbital methods are aimed in developing a theoretical model that gives an alternative choice for experiment by making accurate prediction of chemical systems. *Ab initio* methods can even predict properties of chemical systems that are not suitable to be measured experimentally. They are derived without empirical or experimental factors other than the fundamental physical constants.

II.B.1 The Hartree-Fock theory

The Hartree-Fock (HF) theory plays a crucial role for the development of *ab initio* molecular orbital methods. It introduces in a natural way the concept of molecular orbitals which has become an important conceptual tool to describe qualitatively the electronic structure of complex systems. In addition, it is (often) used as the starting point for more quantitative treatments of many-electron systems, introducing electron correlation.

The HF theory provides approximate GS solution of the non-relativistic electronic Schrödinger equation (Eq.(2-9)) stemming from the application of the variational principle, Eq. (2-11), to the trial function in the form of a single Slater determinant. Such wavefunction satisfies the fundamental property: it is antisymmetric with respect to interchange of the spatial-spin coordinates of any two electrons

$$\Psi\left(\mathbf{x}_1,\cdots,\mathbf{x}_i,\cdots,\mathbf{x}_j,\cdots,\mathbf{x}_N\right)=-\Psi\left(\mathbf{x}_1,\cdots,\mathbf{x}_j,\cdots,\mathbf{x}_i,\cdots,\mathbf{x}_N\right). \qquad (2\text{-}19)$$

Generally, for N electrons and N orthonormal spin orbitals $\phi_i\left(\mathbf{x}\right)$, the Slater determinant is

$$\begin{aligned}\Phi\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\right)&=\frac{1}{\sqrt{N!}}\begin{vmatrix}\phi_1\left(\mathbf{x}_1\right) & \phi_1\left(\mathbf{x}_2\right) & \cdots & \phi_1\left(\mathbf{x}_N\right)\\ \phi_2\left(\mathbf{x}_1\right) & \phi_2\left(\mathbf{x}_2\right) & \cdots & \phi_2\left(\mathbf{x}_N\right)\\ \cdots & \cdots & \cdots & \cdots\\ \phi_N\left(\mathbf{x}_1\right) & \phi_N\left(\mathbf{x}_2\right) & \cdots & \phi_N\left(\mathbf{x}_N\right)\end{vmatrix}\\ &=\frac{1}{\sqrt{N!}}\det\left[\phi_i\left(\mathbf{x}_j\right)\right]_{i,j=1}^{N}=\left(N!\right)^{-1/2}\det\left[\phi_1\left(\mathbf{x}_1\right),\cdots,\phi_N\left(\mathbf{x}_N\right)\right].\end{aligned} \qquad (2\text{-}20)$$

The one-electron spin orbital $\phi_i\left(\mathbf{x}\right)$ is the product of one-electron spatial orbital $\psi_i\left(\mathbf{r}\right)$ and spin function $\sigma_i\left(\kappa\right)$:

$$\phi_i\left(\mathbf{x}\right)=\psi_i\left(\mathbf{r}\right)\sigma_i\left(\kappa\right); \qquad \sigma_i\left(\kappa\right)\in\left\{\alpha\left(\kappa\right),\beta\left(\kappa\right)\right\} \qquad (2\text{-}21)$$

For computational convenience, spin orbitals are usually chosen to be orthonormal

$$\int \phi_i(\mathbf{x})\phi_j(\mathbf{x})d\mathbf{x} = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{2-22}$$

and the orthonormality of spin functions is

$$\langle \alpha | \alpha \rangle = \langle \beta | \beta \rangle = 1; \qquad \langle \alpha | \beta \rangle = 0\,. \tag{2-23}$$

The Slater determinant is normalized, $\langle \Phi | \Phi \rangle = 1$. According to the variational principle, the best spin orbitals are obtained by minimizing the total electronic energy

$$\mathrm{E}_{\mathrm{HF}} = \langle \Phi | \hat{H}_{\mathrm{e}} | \Phi \rangle = \mathrm{E}\left[\Phi\left(N, \{\phi_i\}\right)\right] = \sum_{i=1}^{N} h_{ii} + \frac{1}{2}\sum_{i,j=1}^{N}\left(J_{ij} - K_{ij}\right). \tag{2-24}$$

The one-electron contribution

$$h_{ii} = \langle \phi_i | \hat{h} | \phi_i \rangle \tag{2-25}$$

is due to one-electron operators in Eq.(2-8), where

$$\hat{h}(\mathbf{r}) = -\frac{1}{2}\nabla^2(\mathbf{r}) + v(\mathbf{r}); \qquad v(\mathbf{r}) = -\sum_{A=1}^{P}\frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|}. \tag{2-26}$$

The combination of the coulombic J_{ij} and exchange K_{ij} energy integrals in Eq.(2-24) represents $\langle \Phi | \hat{V}_{\mathrm{ee}} | \Phi \rangle$. Here

$$J_{ij} = \int \phi_i(\mathbf{x}_1)\phi_i^*(\mathbf{x}_1)\frac{1}{r_{12}}\phi_j^*(\mathbf{x}_2)\phi_j(\mathbf{x}_2)d\mathbf{x}_1 d\mathbf{x}_2\,, \tag{2-27}$$

$$K_{ij} = \int \phi_i^*(\mathbf{x}_1)\phi_j(\mathbf{x}_1)\frac{1}{r_{12}}\phi_i(\mathbf{x}_2)\phi_j^*(\mathbf{x}_2)d\mathbf{x}_1 d\mathbf{x}_2\,. \tag{2-28}$$

Note that $J_{ii} = K_{ii}$ (this removes the self-interaction error) and $J_{ij} \geq K_{ij} \geq 0$. The minimization of E_{HF}, Eq.(2-24), under the constraint in Eq.(2-22) means that the Lagrange function L should be stationary with respect to an orbital variation ($\varepsilon_{ij} = \varepsilon_{ij}^*$ are Lagrange multipliers), $L = E_{\mathrm{HF}} - \sum_{i,j=1}^{N}\varepsilon_{ij}\left(\langle \phi_i | \phi_j \rangle - \delta_{ij}\right)$. This yields the Hartree–Fock equations of the form[3-5]

$$\hat{F}\phi_i = \sum_{j=1}^{N}\varepsilon_{ij}\phi_j\ , \tag{2-29}$$

or a simplified form, usually called canonical form, after appropriate unitary transformation, $\phi_i' = \sum_j U_{ij}\phi_i$ (this also preserves the orthonormality of spin orbitals),

$$\hat{F}\phi_i' = \varepsilon_i\phi_i', \qquad i = 1,\cdots,N; \quad \varepsilon_i \le \varepsilon_{i+1}. \tag{2-30}$$

The prime in the orbital can be omitted because the Fock operator remains unchanged after unitary transformation. The Fock operator $\hat{F}$ is an effective one-electron operator,

$$\hat{F} = \hat{h} + \hat{j} - \hat{k}, \tag{2-31}$$

where $\hat{h}$ is given in Eq.(2-26), the Coulomb and exchange operators are defined as

$$\hat{j}(\mathbf{x}_1)f(\mathbf{x}_1) = \sum_{i=1}^{N}\int\left|\phi_i(\mathbf{x}_2)\right|^2 \frac{1}{r_{12}} f(\mathbf{x}_1)\,d\mathbf{x}_2, \tag{2-32}$$

$$\hat{k}(\mathbf{x}_1)f(\mathbf{x}_1) = \sum_{i=1}^{N}\int \phi_i^*(\mathbf{x}_2)f(\mathbf{x}_2)\frac{1}{r_{12}}\phi_i\,(\mathbf{x}_1)\,d\mathbf{x}_2, \tag{2-33}$$

with $f(\mathbf{x}_1)$ an arbitrary function. Note that $\hat{j}(\mathbf{x}_1)$ is a multiplicative operator (electrostatic potential), while $\hat{k}(\mathbf{x}_1)$ is an integral operator (the Fock exchange potential operator). In Eq.(2-31) $\hat{h}$ describes the motion of single electron in the presence of the external potential $v(\mathbf{r})$, while $\hat{j} - \hat{k}$ describes the average repulsive potential energy experienced by this electron due to (presence of) remaining $N-1$ electrons.

The expectation value of $\hat{F}$ is the orbital energy ε_i of the molecular spin orbital,

$$\varepsilon_i = \langle\phi_i|\hat{F}|\phi_i\rangle = h_{ii} + \sum_{j=1}^{N}\left[J_{ij} - K_{ij}\right]. \tag{2-34}$$

If we substitute h_{ii} from Eq.(2-34) into Eq.(2-24), the total electronic energy becomes

$$E_{\mathrm{HF}} = \sum_{i}^{N}\varepsilon_i - \frac{1}{2}\sum_{i,j=1}^{N}\left[J_{ij} - K_{ij}\right], \tag{2-35}$$

this shows that the total electronic energy is not the sum of the orbital energies of electrons.

Depending on the restriction on spin orbitals used, HF can be classified as restricted (RHF), unrestricted HF (UHF) and restricted-open shell one (ROHF). Applied to closed-shell systems, when all electrons are paired (two electrons of opposite spin per one occupied spatial orbital) the formalism is known as the RHF. Using the basis set representation for the molecular orbitals, Eq.(2-15), a generalized matrix eigenvalue type equations for RHF in the matrix notation is

$$\mathbf{FC} = \mathbf{SC}\boldsymbol{\varepsilon}, \tag{2-36}$$

where **F**, **S**, **ε**, and **C** are the $K \times K$ matrices representing the Fock, overlap, orbital energy and coefficients, respectively.

The UHF method is used for open-shell molecules where the numbers of electrons of each spin are not equal. Different spatial molecular orbitals for the α and β electrons are applied. In this case, Eq.(2-36) is replaced by a pair of coupled Roothaan equations separately for α and β electrons. While computational efficiency is the strong side of UHF, spin contamination is a problem. This is due to the fact that the single Slater determinant of UHF is not an eigenfunction of the squared total spin operator $\hat{S}^2$. When the errors introduced by spin contamination are high and unacceptable, the restricted open-shell HF (ROHF) is the best choice[10] (the spin contamination is eliminated by construction of this method).

Since originally the spatial orbitals were expanded in K basis functions, the solution of the UHF equations yields $2K$ molecular spin orbitals (N MSOs are occupied and $2K$ - N are unoccupied). The latter are also called virtual orbitals. Improving the basis set, that is increasing the number of basis functions K, results in the decrease of the HF energy, because of the better flexibility of the wave function. For $K \to \infty$, one obtains the best possible energy within the Hartree-Fock approximation, i.e. the Hartree-Fock limit.

Various mathematical and computational efforts have been done to reach the HF limit with no additional approximations. HF limit is the lowest energy that can be obtained from a single Slater determinant. The energy error associated with the HF approximation for a given system is known as the electron correlation energy E_c. It is defined by

$$E_\mathrm{c} = E_0 - E_\mathrm{HF} \le 0 \tag{2-37}$$

with E_0 being the true GS energy while E_HF is Hartree-Fock energy at HF limit.

II.B.2 Post Hartree-Fock methods

While in the HF method the motion of electrons with the same spin is correlated in Hartree-Fock through the exchange operator, one of its disadvantages is the lack of a correlation in the motion of electrons with opposite spins. A number of methods is available to account for the effects of correlation. We will focus exclusively here on the so-called *single-reference methods*, starting from a single HF determinant. Some of the most common post-HF methods include the configuration interaction, coupled-cluster and Møller-Plesset perturbation theory.

A first method is the configuration interaction (CI). Here, a linear combination Ψ of all possible N-electron Slater determinants formed from a complete set of spin orbitals is used, which can be shown to be equal to the exact trial wavefunction for an N-electron system. In the CI method, the "basis set" of Slater determinants is generated first by computing the HF Φ_0 as usual, which also generates a lot virtual MOs, and, next, generating other determinants by replacing one or more occupied MOs with virtual ones. In general

$$\left|\Psi\right\rangle = c_0\left|\Phi_0\right\rangle + \sum_{i}^{occ}\sum_{a}^{vir} c_i^a \left|\Phi_i^a\right\rangle + \sum_{i<j}^{occ}\sum_{a<b}^{vir} c_{ij}^{ab}\left|\Phi_{ij}^{ab}\right\rangle + \cdots, \tag{2-38}$$

where Φ_i^a denotes the modified Slater determinant due to single excitation (in which the occupied MO ψ_i is replaced by the virtual MO ψ_a); Φ_{ij}^{ab} denotes a Slater determinant arising due to double excitation: replacement of the *i* and *j* occupied MOs by the *a* and *b* virtual MOs. The first term in Eq.(2-38) is the HF Slater determinant. The coefficients $\{c\}$ are determined variationally (under the normalization constraint) using the CI secular equation.[13,14]

In the basis set limit of the starting HF results, a full CI (FCI) calculation, (infinite series in Eq.(2-38)), gives the exact solution to the electronic Schrodinger equation, Eq. (2-9). A FCI calculation within a given (limited) basis provides a variational energy and is size consistent (the energy of two infinitely separated molecules will be the same as the sum of the energies obtained from two individual calculations at the same theory and basis set level). However, since the number of excited Slater determinants grow factorially with the system size, therefore in practice, Eq.(2-38) needs to be truncated at certain level of excitation. It can include only all single excitations (CIS), single + double excitations (CISD), or single + double + triple excitations (CISDT). Unlike FCI, a truncated CI is not invariant to a unitary transformation of the molecular orbitals and is size inconsistent[15,16].

In multiconfigurational self-consistent field (MCSCF) theory[12] the trial wavefunction represents suitably truncated CI expansion, Eq.(2-38) (a linear combination of selected configurations). The optimum wavefunction is constructed by variationally optimizing both the coefficients at configurations (Slater determinants) and expansion coefficients of MOs in their basis. This theory allows for describing the entire correlation energy (dynamic and nondynamic correlations). Many MCSCF approaches can generate very accurate wave functions, however they are computationally very intensive and are not size extensive.

Coupled-cluster (CC) theory is based on the idea of describing the electron correlation in terms of interacting clusters of electrons and aimed to include all corrections of a given type to infinite order. The formalism is based on the exponential wavefunction ansatz. For an *N*- electron wave function Ψ , the CC form is

$$\left|\Psi\right\rangle = \mathrm{e}^{\hat{T}}\left|\Phi_0\right\rangle, \tag{2-39}$$

where the excitation operator

$$e^{\hat{T}} = 1 + \hat{T} + \frac{\hat{T}^2}{2!} + \frac{\hat{T}^3}{3!} + \cdots = \sum_k^{\infty} \frac{\hat{T}^k}{k!}, \tag{2-40}$$

uses the cluster operator, $\hat{T} = \hat{T}_1 + \hat{T}_2 + \cdots + \hat{T}_N$ with $\hat{T}_i$ generated from *i*-fold excitations. Each $\hat{T}_i$ operator is the sum of the operators with particular *i*-tuple excitation multiplied by its amplitude *t*. For example, the cluster operator of all single excitations, $\hat{T}_1$ and the operator of all double excitations $\hat{T}_2$ are

$$\hat{T}_1\Phi_0 = \sum_i \sum_a t_i^a \Phi_i^a, \quad \hat{T}_2\Phi_0 = \sum_{j>i} \sum_{b>a} t_{ij}^{ab} \Phi_{ij}^{ab} \tag{2-41}$$

whereas $\hat{T}_2^2\Phi_0 = \sum_{j>i} \sum_{k>l} \sum_{b>a} \sum_{d>c} t_{ij}^{ab} t_{kl}^{cd} \Phi_{ijkl}^{abcd}$. The Schrödinger equation with Ψ in the form Eq. (2-39) (truncated at some excitation level) is solved approximately by means of projection. The set of amplitudes $\{t\}$ is determined. As opposed to CID, the CCD (which is simply the coupled cluster $\hat{T}$ with only the double-excitation operator $\hat{T}_2$) includes the quadruple, hexuple, etc., excitations (up to *N*-tuples for a system with *N* electrons). The most robust and most commonly used CC method is the CCSD(T) (coupled cluster for singles/doubles/selected triples coupling terms). CC method ensures size consistency.

The last method in this section is the Møller-Plesset (MP) perturbation theory. This method uses a standard mathematical technique in physics to compute corrections to a reference state (in this case RHF).[3,4,6] Generally, for a given reference Hamiltonian (zero order, $\hat{H}_0$) operator and a perturbed one, $\hat{H}'$, the basic theory uses Hamiltonian

$$\hat{H} = \hat{H}_0 + \lambda \hat{H}' \tag{2-42}$$

where λ is a (variable) parameter that measures the strength of the perturbation. For the zero-order wavefunction the HF determinant is used and the zero-order energy is taken as the sum of the MO energies. The first order-correction includes some the electron-electron repulsion, so the HF energy is the zero-order energy corrected through first-order. The second order correction (MP2) includes some double excitations, the third order correction includes more double excitations than MP2, and MP4 includes single, double, triple, and some quadruple excitations. MP2, for instance typically recovers 80-90 % of the correlation energy. The MP methods are important to produce size consistent and size-extensive results, i.e., where the correlation energy per particle scales linearly as the number of particles increase, but in the case of open-shell systems often suffers from effects of spin-contamination.

For the single-determinant-based methods with a medium-sized basis set, the order of accuracy is HF<< MP2 < CISD = MP4 (SDQ)~ CCSD < MP4 < CCSD(T) < FCI while in terms of computational cost: HF (M^4) < MP2 (M^5) < CCD~CCSD (M^6) < MP4, CCSD(T) (M^7) < FCI. For many applications the sufficient accuracy may be obtained with CCSD. The more accurate (and more expensive) CCSD(T) is often called *the gold standard of quantum chemistry* for its excellent compromise between the accuracy and the computational cost for the molecules near equilibrium geometries, as long as a single HF determinant can be used as a reference for the cluster expansion.

II.C Density functional theory

As it is already discussed in the previous sections, Hartree-Fock method lacks treatment of electron-correlation while post Hartree-Fock methods consume high computational time. So, it is not surprising to look for another method that does not directly depend on the size of the system (atom or molecule) of interest. Introducing an electron density as a basic variable (instead of a wave function) is an alternative to *ab initio* methods[17-19]. The former is straightforwardly linked to the latter as

$$\begin{aligned}\rho(\mathbf{x}_1) &= N\int \Psi(\mathbf{x}_1,\cdots,\mathbf{x}_N)\Psi^*(\mathbf{x}_1,\cdots,\mathbf{x}_N)\,d\mathbf{x}_2\cdots d\mathbf{x}_N \\ &= N\sum_{\kappa_2,\kappa_3,\cdots,\kappa_N}\int\left|\Psi(\mathbf{r}_1,\kappa_1,\cdots,\mathbf{r}_N,\kappa_N)\right|^2 d\mathbf{r}_2 d\mathbf{r}_3\cdots d\mathbf{r}_N\end{aligned}. \tag{2-43}$$

In the spinless approach, $\rho_{\mathrm{N}}(\mathbf{r}) = \sum_{\kappa} \rho(\mathbf{r},\kappa)$ is used, with the subscript N often suppressed. The electron density is therefore a function of a three-dimensional variable and integrates to the total number of electrons of the system:

$$\int \rho(\mathbf{x}_1)\, d\mathbf{x}_1 = N\,. \qquad (2\text{-}44)$$

The ultimate goal of this method is to replace the complicated *N*-electron wavefunction and the associated Schrödinger equation by the much simpler electron density and the associated scheme to determine it.

An early attempt to construct the density functional theory (DFT) was the Thomas-Fermi theory, where the kinetic energy of an atom is expressed in terms of electron density functional while the nuclear-electron and electron-electron contributions are treated in a classical way. The kinetic energy is approximated (from statistical model) by

$$T_{\mathrm{TF}}[\rho] = C_{\mathrm{F}} \int \rho^{5/3}(\mathbf{r})\, d\mathbf{r}; \qquad C_{\mathrm{F}} = \frac{3}{10}\left(3\pi^2\right)^{2/3}, \qquad (2\text{-}45)$$

the classical nuclear-electron energy is

$$E_{\mathrm{ne}} = \int v(\mathbf{r})\rho(\mathbf{r}) d\mathbf{r}\,, \qquad (2\text{-}46)$$

see Eq.(2-26) for $v(\mathbf{r})$, and the classical electron-electron interaction is

$$J[\rho] = \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2\,. \qquad (2\text{-}47)$$

The Thomas-Fermi theory, in general, minimizes the total energy

$$E_{\mathrm{TF}}[\rho] = T_{\mathrm{TF}}[\rho] + E_{\mathrm{ne}}[\rho] + J[\rho]\,, \qquad (2\text{-}48)$$

over $\rho(\mathbf{r})$ under the constraint $\int \rho(\mathbf{r}) d\mathbf{r} = N$. Due to assumptions, $E_{\mathrm{TF}}[\rho]$ is a very crude approximation to $E[\rho]$. This theory gives only a very coarse approximation. It is exact only in the limit of an infinite nuclear charge.

II.C.1 The Hohenberg-Kohn theorems

The major theoretical pillars of modern DFT was embarked in 1964 by the pioneer works of Hohenberg and Kohn. The concepts of their theory originate from two theorems[17].

The first theorem states that "*the electron density* $\rho(\mathbf{r})$ *determines the external potential* $v(\mathbf{r})$ *within a trivial additive constant*". The proof of the first HK theorem is given by *a reductio ad absurdum*, starting from the hypothesis that two external potentials $v(\mathbf{r})$and $v'(\mathbf{r})$ are linked to the same $\rho(\mathbf{r})$. Since N and $v(\mathbf{r})$ determine the molecular Hamiltonian, the full many-particle ground state Ψ is a functional of $\rho(\mathbf{r})$. In more elaborated way, the electron ground state density, $\rho_{\text{gs}}(\mathbf{r})$ uniquely determines the properties of the system.

The energy of the system as a functional of electron density is written as[19,20],

$$E[\rho] = E_{\text{HK}}[\rho] = \langle \Psi[\rho] | \hat{H} | \Psi[\rho] \rangle = \int v(\mathbf{r}) \rho(\mathbf{r}) d\mathbf{r} + F_{\text{HK}}[\rho], \tag{2-49}$$

where, $F_{\text{HK}}[\rho]$ is the Hohenberg-Kohn universal functional

$$F_{\text{HK}}[\rho] = \langle \Psi[\rho] | \hat{T} + \hat{V}_{\text{ee}} | \Psi[\rho] \rangle = T[\rho] + V_{\text{ee}}[\rho]. \tag{2-50}$$

Here $T[\rho]$ represents the kinetic energy functional, $V_{\text{ee}}[\rho]$—the electron-electron interaction energy functional. It should be noted that $F_{\text{HK}}[\rho]$ is independent of the external potential.

The second Hohenberg-Kohn theorem establishes the energy variational principle, stating that "*the exact ground state electron density minimizes the exact energy functional*". It states that the energy functional $E_{\text{HK}}[\rho]$ of Eq.(2-49) satisfies $E_{\text{HK}}[\rho] \geq E_{\text{HK}}[\rho_0] = E_0$ i.e. for any trial density the corresponding energy is an upper limit to the ground state electronic energy E_0 and to attain the minimum value of the energy with respect to all allowed densities, the input density have to be the same as the true ground state density ρ_0. The necessary condition for the minimum energy is

$$\frac{\delta E_{\text{HK}}[\rho]}{\delta \rho(\mathbf{r})} = \frac{\delta F_{\text{HK}}[\rho]}{\delta \rho(\mathbf{r})} + \text{v}(\mathbf{r}) = \text{const}. \tag{2-51}$$

This can be considered as the DFT analogue of the Schrödinger equation.

II.C.2 The Kohn-Sham equations

The full practical implementation of DFT is based on a fictitious system of non-interacting electrons subject to such external potential (instead of interaction with the nuclei) that preserves the exact density distribution of the true system, so called the Kohn-Sham method[18]. The wavefunction of this system is a Slater determinant of N spin orbitals $\{\phi_i\}$. The exact kinetic energy $T_\mathrm{s}[\rho]$ of the noninteracting reference system is

$$T_\mathrm{s}[\rho] = -\frac{1}{2}\sum_i^N \langle \phi_i | \nabla^2 | \phi_i \rangle, \tag{2-52}$$

and $T_\mathrm{s}[\rho] \neq T[\rho]$ for the same density. The one-electron density is

$$\rho(\mathbf{r}) = \sum_\kappa \sum_{i=1}^N |\phi_i(\mathbf{r},\kappa)|^2, \tag{2-53}$$

Which, by constructing, equals the electron density of the real system. The universal functional F_HK, Eq.(2-50), can be rewritten for the Kohn-Sham DFT as

$$F[\rho] = T_s[\rho] + J[\rho] + E_\mathrm{xc}[\rho] = F_\mathrm{HK}[\rho], \tag{2-54}$$

where $E_\mathrm{xc}[\rho]$ is the exchange-correlation functional

$$E_\mathrm{xc}[\rho] = (T[\rho] - T_\mathrm{s}[\rho]) + (V_\mathrm{ee}[\rho] - J[\rho]) \tag{2-55}$$

which contains the kinetic correlation energy as well as the non-classical part of the V_ee energy. The explicit form of this functional is not known, but various approximate density functionals for it have been proposed.

Minimization over the density of the total energy functional (see Eq.(2-49)) in the KS form

$$E[\rho] = \int v(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + F[\rho] = T_\mathrm{s}[\rho] + V_\mathrm{eff}[\rho] \tag{2-56}$$

where (see Eq.(2-54))

$$V_\mathrm{eff}[\rho] = \int v(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + J[\rho] + E_\mathrm{xc}[\rho], \tag{2-57}$$

can be replaced by the minimization over orbitals $\{\phi_i\}$, due to $T_\mathrm{s}[\rho]$ and ρ depending on orbitals, Eqs.(2-52) and (2-53). This leads to the KS equations

$$\left(-\frac{1}{2}\nabla^2 + \mathrm{v}_\mathrm{eff}(\mathbf{r})\right)\phi_i(\mathbf{r},\kappa) = \varepsilon_i\phi_i(\mathbf{r},\kappa), \qquad \varepsilon_i \le \varepsilon_{i+1}, \quad i = 1,2,\cdots,N, \tag{2-58}$$

similar to the HF Eq.(2-30), but with a different potential

$$v_\mathrm{eff}(\mathbf{r}) = \frac{\delta V_\mathrm{eff}[\rho]}{\delta\rho(\mathbf{r})} = v(\mathbf{r}) + v_\mathrm{H}(\mathbf{r}) + v_\mathrm{xc}(\mathbf{r}), \tag{2-59}$$

where

$$v_{\mathrm{H}}(\mathbf{r}) = \frac{\delta J[\rho]}{\delta\rho(\mathbf{r})} = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' \tag{2-60}$$

$$v_{\mathrm{xc}}(\mathbf{r}) = \frac{\delta E_{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})}. \tag{2-61}$$

In fact, the effective HF potential contains the integral operator $-\hat{k}$ instead of the local v_{xc} in the KS potential, Eq.(2-59). Since v_{eff} is a functional of ρ (see v_{H} and v_{xc}), KS equations are to be solved self-consistently.

In general, DFT is an exact theory (in contrast to HF which is approximate theory). However, in practice, approximate functionals for $E_{\mathrm{xc}}[\rho]$ are used (density functional approximations, DFAs), therefore the DFA results cannot be exact, their accuracy depends on a quality of approximation for $E_{\mathrm{xc}}[\rho]$. Due to the fact that the approximate $E_{\mathrm{xc}}[\rho]$ does not cancel exactly the self-interaction terms in $J[\rho]$, the DFA is not self-interaction free.

II.C.3 Exchange-correlation functionals

The challenging task in DFT is the development of an accurate functional for expressing the exchange-correlation energy, E_{xc} in terms of density ρ. Although no explicit form is available for the exact $E_{\mathrm{xc}}[\rho]$, much is known about the "proper" way in which approximations for this energy term should be constructed. One of the main challenges for DFT is to keep, as its cornerstone, some element of simplicity[21]. Unfortunately, a simple density functional adopted from the solid state physics, the LDA, does not perform well in many areas of chemistry. The introduction of the density gradient into the form of the exchange-correlation functional (the GGA or meta-GGA), allows to receive satisfactory results in the chemical applications. The next major advances came with the inclusion of a fraction of the HF exchange in the E_{xc} functional (hybrid functionals). One of the advanced developments in functionals is due to the range separation. The idea is to separate the electron-electron interaction into two parts, the long-range one and the short-range one, and then to treat these parts with different functionals[22-30] (see Ref. [21] for review).

Many density functionals are available at present. There is no consensus on which functional is the best overall, and the fact that B3LYP is still a widely used functional, even though it was developed in 1993, is very telling. Some examples of functionals are presented below.

The first simple and practical approximation for $E_{\mathrm{xc}}\left[\rho\right]$ is the so-called local density approximation (LDA)[31]

$$E_{\mathrm{xc}}^{\mathrm{LDA}}\left[\rho\right]=\int\rho\left(\mathbf{r}\right)\varepsilon_{\mathrm{xc}}\left(\rho\left(\mathbf{r}\right)\right)d\mathbf{r}\,,\tag{2-62}$$

which enables DFT to be popular. The $\varepsilon_{\mathrm{xc}}\left(\rho\right)$ is the exchange-correlation energy per particle (electron) of a uniform electron gas of density ρ. In practice, the exchange $E_{\mathrm{x}}^{\mathrm{LDA}}$ and correlation $E_{\mathrm{c}}^{\mathrm{LDA}}$ energies are calculated separately. The former one is analytically known exactly,

$$E_{\mathrm{x}}^{\mathrm{LDA}}\left[\rho\right]=-\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int\rho^{4/3}\left(\mathbf{r}\right)d\mathbf{r}\,,\tag{2-63}$$

while the later (difficult to find analytically) is obtained by fitting to the Monte-Carlo results for energy of the many-particle free electron gas[32-34]. LDA is valid only for slowly varying densities. So, an improvement of LDA is necessary. One type of improvement is the general gradient approximation (GGA)[35] having a functional form

$$E_{\mathrm{xc}}^{\mathrm{GGA}}\left[\rho\right]=\int f\left(\rho\left(\mathbf{r}\right),\nabla\rho\left(\mathbf{r}\right)\right)d\mathbf{r}\,.\tag{2-64}$$

This is semi-local functional that depends on both density and its gradient – for the correction of density inhomogeneity. The exchange energy in GGA formalism is

$$E_{\mathrm{x}}^{\mathrm{GGA}}\left[\rho\right]=\int\rho\left(\mathbf{r}\right)\varepsilon_{\mathrm{x}}\left(\rho\left(\mathbf{r}\right)\right)f_{\mathrm{x}}^{\mathrm{GGA}}\left(s\left(\mathbf{r}\right)\right)d\mathbf{r}\,,\tag{2-65}$$

with $f_{\mathrm{x}}^{\mathrm{GGA}}$ being the exchange correction term to the LDA, a function of the dimensionless reduced gradient

$$s\left(\mathbf{r}\right)=\frac{\left|\nabla\rho\left(\mathbf{r}\right)\right|}{2\left(3\pi^{2}\right)^{1/3}\rho^{4/3}\left(\mathbf{r}\right)}\,.\tag{2-66}$$

One popular example of a GGA functional is due to the Perdew, Burke and Ernzerhof (PBE)[36,37] of the form

$$f_{\mathrm{x}}^{\mathrm{PBE}}\left(s\right)=1+a-\frac{a}{\left(1+bs^{2}/a\right)}\,,\tag{2-67}$$

a and *b* are parameters obtained from non-empirical physical constraints. If $f_{\mathrm{x}}^{\mathrm{GGA}}(s)$ is unity, the GGA, Eq. (2-65) is reduced to LDA, Eq.(2-63). The gradient corrected correlation energy is expressed as complex function of s.

Another well known exchange correlation functionals are the hybrid functionals that are mixtures of the Hartree-Fock exchange with DFT exchange-correlation. Among them the most popular is the Becke type three parameter Lee-Yang-Parr (B3LYP)[38] one of the form

$$E_{\mathrm{xc}}^{\mathrm{B3LYP}} = (1-a)E_{\mathrm{x}}^{\mathrm{LSDA}} + aE_{\mathrm{x}}^{\mathrm{HF}} + b\Delta E_{\mathrm{x}}^{\mathrm{B88}} + (1-c)E_{\mathrm{c}}^{\mathrm{LSDA}} + cE_{\mathrm{c}}^{\mathrm{LYP}}. \tag{2-68}$$

The coefficients *a*, *b* and *c* are 0.20, 0.72, and 0.81, respectively. Here $\Delta E_{\mathrm{x}}^{\mathrm{B}}$ is Becke's 1988 functional, which includes the Slater exchange along with corrections involving the gradient of the density[39], $E_{\mathrm{c}}^{\mathrm{LYP}}$ is the correlation functional of Lee, Yang, and Parr, which includes both local and non-local terms[40]. The $E_{\mathrm{x}}^{\mathrm{LSDA}}$ and $E_{\mathrm{x}}^{\mathrm{HF}}$ represent local spin density approximation and Hartree-Fock expression, respectively.

The example of the functional which includes long-range corrections using the Coulomb-attenuating method is rCAM-B3LYP[41] functional (a long-range corrected functional with improved description of systems with fractional numbers of electrons)

$$E_{\mathrm{xc}}^{\mathrm{rCAM\text{-}B3LYP}} = \begin{pmatrix} \alpha E_{\mathrm{x}}^{\mathrm{HF}} + (1-\alpha)E_{\mathrm{x}}^{\mathrm{LSDA}} + c^{\mathrm{B88}}E_{\mathrm{x}}^{\mathrm{B88}} + \beta\left(E_{\mathrm{x}}^{\mathrm{LR,HF}} - E_{\mathrm{x}}^{\mathrm{LR,B88}}\right) \\ +dE_{\mathrm{c}}^{\mathrm{LYP}} + (1-d)E_{\mathrm{c}}^{\mathrm{VWN}} \end{pmatrix} \tag{2-69}$$

Here $E_{\mathrm{x}}^{\mathrm{LR,HF}}$ and $E_{\mathrm{x}}^{\mathrm{LR,HF}}$ are the long-range HF exchange and the long-range DFA exchange functionals. They depend on the parameter which determines the splitting into short- and long-range parts.

II.D Density matrices

An ensemble description of quantum states becomes necessary when, e.g. the state cannot be represented by a linear combination of eigenstates of a particular Hamiltonian. A state is said to be *pure* if it is described by some state vector $|\Psi\rangle$ (wavefunction) of the Hilbert space, being normalized$\langle\Psi|\Psi\rangle = 1$, *mixed* (ensemble) if such description is impossible. For example, a macroscopic system in a thermal equilibrium represents some mixed state. A system in a mixed state can be characterized by a probability distribution over all the accessible pure states and is described by the *density matrix operator*.

In this section, the density matrix operator will be recalled. The density matrix formalism of increasing complexity, starting with the pure state, continuing with ensemble state in the Hilbert space, and ending with the ensemble state in the Fock space is given.

II.D.1 Pure state

An N-electron Hilbert space $\mathcal{Y}^N$ is defined as the following set of state vectors

$$\mathcal{Y}^N = \left\{ \left| \Psi^N \right\rangle \,\middle|\, \left\langle \Psi^N \middle| \Psi^N \right\rangle = 1, \hat{\mathcal{N}} \left| \Psi^N \right\rangle = N \left| \Psi^N \right\rangle \right\}, \tag{2-70}$$

where $\hat{\mathcal{N}}$ is the particle-number operator. A state vector $\left|\Psi\right\rangle$ (the superscript N is suppressed when obvious) in the basis $\left|\mathbf{x}_1,\cdots,\mathbf{x}_N\right)$ equals N-electron wavefunction in the configurational space

$$\left(\mathbf{x}_1,\cdots,\mathbf{x}_N \middle| \Psi\right\rangle = \Psi\left(\mathbf{x}_1,\cdots,\mathbf{x}_N\right), \tag{2-71}$$

which is antisymmetric (see Eq.(2-19)).

The N-body density matrix (N-DM), $\gamma_N\left(\mathbf{x}_1,\cdots,\mathbf{x}_N;\mathbf{x}'_1,\cdots,\mathbf{x}'_N\right)$, for a pure state $\left|\Psi\right\rangle$ is the product of N-electron wavefunctions

$$\gamma_N\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N;\mathbf{x}'_1,\mathbf{x}'_2,\cdots,\mathbf{x}'_N\right) = \Psi\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\right)\Psi^*\left(\mathbf{x}'_1,\mathbf{x}'_2,\cdots,\mathbf{x}'_N\right) \equiv \gamma_N\left[\Psi^N\right]. \tag{2-72}$$

The primed $\mathbf{x}_i$ is used to distinguish it from the corresponding unprimed ones. The DM γ_N, Eq.(2-72), is the kernel of the density matrix operator,

$$\hat{\gamma}_N = \left|\Psi\right\rangle\left\langle\Psi\right|, \tag{2-73}$$

acting in N-electron Hilbert space. It fulfills properties such as idempotency,

$$\hat{\gamma}_N \cdot \hat{\gamma}_N = \hat{\gamma}_N, \tag{2-74}$$

unit-trace one,

$$\mathrm{tr}\left(\hat{\gamma}_N\right) = 1, \tag{2-75}$$

Hermiticity, and positive semi-definiteness. The satisfaction of Eq.(2-74) is necessary and sufficient condition for $\hat{\gamma}_N$ to be a pure state DM.

The expectation (mean) value of an observable operator $\hat{A}$ in the pure state $|\Psi\rangle$ can be written as the trace of this operator with $\hat{\gamma}_N[\Psi]$:

$$\begin{aligned} A &= \left\langle \Psi \middle| \hat{A} \middle| \Psi \right\rangle = \operatorname{tr} \hat{\gamma}_N \hat{A} = \operatorname{tr} \hat{A} \hat{\gamma}_N \\ &= \int \Psi^* \left(\mathbf{x}_1, \cdots, \mathbf{x}_N \right) \hat{A} \Psi \left(\mathbf{x}_1, \cdots, \mathbf{x}_N \right) d\mathbf{x}_1 ... d\mathbf{x}_N \end{aligned} \tag{2-76}$$

In view of Eq.(2-76), $\hat{\gamma}_N$ carries the same information as the N-electron wavefunction $|\Psi\rangle$. Note that while $|\Psi\rangle$ is defined only up to an arbitrary phase factor, $\hat{\gamma}_N$ for the state $|\Psi\rangle$ is unique.[39]

From an application point of view, basic physical quantities of a system such as energy (expectation value of the Hamiltonian operator) are expressed in terms of one- and two-electron density operators (as two degrees of freedom are sufficient to describe operators of the system). The N-DM can be reduced into p-order (p-body) density matrix (p-DM), γ_p^N, (where $1 \leq p < N$) as[3,32,42,43]

$$\begin{aligned} &\gamma_p^N \left(\mathbf{x}_1, \cdots, \mathbf{x}_p; \mathbf{x}_1', \cdots, \mathbf{x}_p' \right) \\ &\quad = \binom{N}{p} \int \gamma_N \left(\mathbf{x}_1, \cdots, \mathbf{x}_N; \mathbf{x}_1', \cdots, \mathbf{x}_p', \mathbf{x}_{p+1}, \cdots, \mathbf{x}_N \right) d\mathbf{x}_{p+1} \cdots d\mathbf{x}_N \end{aligned} \tag{2-77}$$

where the factor in front of the integral is the binomial coefficient equal to $N! / \left(p! (N-p)! \right)$. The $\int d\mathbf{x}_i$ means $\sum_{\kappa_i} \int d\mathbf{r}_i$, the integration runs over the whole space of variable $\mathbf{r}_i$ while the summation runs over two values of the spin variable κ_i. In particular, the reduced 2-DM is

$$\begin{aligned} &\gamma_2^N \left(\mathbf{x}_1, \mathbf{x}_2; \mathbf{x}_1', \mathbf{x}_2' \right) \\ &\quad = \frac{N(N-1)}{2} \int \Psi \left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \cdots, \mathbf{x}_N \right) \Psi^* \left(\mathbf{x}_1', \mathbf{x}_2', \mathbf{x}_3 \cdots, \mathbf{x}_N \right) d\mathbf{x}_3 \cdots d\mathbf{x}_N \equiv \gamma_2 \left[\Psi^N \right]. \end{aligned} \tag{2-78}$$

The reduced 1-DM is

$$\gamma_1^N \left(\mathbf{x}_1; \mathbf{x}_1' \right) = N \int \Psi \left(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N \right) \Psi^* \left(\mathbf{x}_1', \mathbf{x}_2, \cdots, \mathbf{x}_N \right) d\mathbf{x}_2 \cdots d\mathbf{x}_N \equiv \gamma_1 \left[\Psi^N \right]. \tag{2-79}$$

The traces (the integrals of the diagonal elements) of $\gamma_1 \left(\mathbf{x}_1; \mathbf{x}_1' \right)$ and $\gamma_2 \left(\mathbf{x}_1, \mathbf{x}_2; \mathbf{x}_1', \mathbf{x}_2' \right)$,

$$\operatorname{tr} \hat{\gamma}_1^N \equiv \int \gamma_1^N \left(\mathbf{x}_1; \mathbf{x}_1 \right) d\mathbf{x}_1 = N, \tag{2-80}$$

$$\mathrm{tr}\ \hat{\gamma}_2^N \equiv \int \gamma_2^N\left(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1,\mathbf{x}_2\right)d\mathbf{x}_1 d\mathbf{x}_2 = \frac{N\left(N-1\right)}{2}, \tag{2-81}$$

equal the electron number and the electron-pair number, respectively.

The γ_1^N and γ_2^N are related as

$$\gamma_1^N\left(\mathbf{x}_1;\mathbf{x}_1'\right) = \frac{2}{N-1}\int \gamma_2^N\left(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1',\mathbf{x}_2\right)d\mathbf{x}_2\ . \tag{2-82}$$

Density matrices possess properties of Hermiticity

$$\gamma_1\left(\mathbf{x}_1;\mathbf{x}_1'\right) = \gamma_1^*\left(\mathbf{x}_1';\mathbf{x}_1\right), \qquad \gamma_2\left(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1',\mathbf{x}_2'\right) = \gamma_2^*\left(\mathbf{x}_1',\mathbf{x}_2';\mathbf{x}_1,\mathbf{x}_2\right), \tag{2-83}$$

and antisymmetry

$$\begin{aligned}\gamma_2\left(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1',\mathbf{x}_2'\right) &= -\gamma_2\left(\mathbf{x}_2,\mathbf{x}_1;\mathbf{x}_1',\mathbf{x}_2'\right)\\ &= -\gamma_2\left(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_2',\mathbf{x}_1'\right) = \gamma_2\left(\mathbf{x}_2,\mathbf{x}_1;\mathbf{x}_2',\mathbf{x}_1'\right).\end{aligned} \tag{2-84}$$

The reduction of *N*-DM into 1-DM and 2-DM is indeed essential, since 1-DM depends only on two space-spin variables while 2-DM depends on four ones, regardless of how large the system size is (how many electrons it has). In general, density matrices γ_1 and γ_2 are used for calculation of expectation values of arbitrary one-body and two-body operators such as kinetic energy and electron-electron interaction energy, respectively.

When the expectation value of a multiplicative operator is calculated as the operator trace, Eq.(2-76), only the knowledge of diagonal parts (by setting $\mathbf{x}_i' = \mathbf{x}_i$ for all *i*) of the reduced 1-DM and 2-DM is needed: the spinor electron density, $\rho\left(\mathbf{x}\right)$, (the diagonal part of the 1-DM), see Eq. (2-43)

$$\rho\left(\mathbf{x}\right) = \rho\left(\mathbf{r},\kappa;\left[\Psi\right]\right) \equiv \gamma_1\left(\mathbf{x};\mathbf{x};\left[\Psi\right]\right), \tag{2-85}$$

and spinor pair density, $g\left(\mathbf{x}_1,\mathbf{x}_2\right)$,(the diagonal part of the 2-DM)

$$g\left(\mathbf{x}_1,\mathbf{x}_2\right) = g\left(\mathbf{r}_1,\kappa_1,\mathbf{r}_2,\kappa_2;\left[\Psi\right]\right) \equiv \gamma_2\left(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1,\mathbf{x}_2;\left[\Psi\right]\right), \tag{2-86}$$

which is the probability density to find one electron at $\mathbf{x}_1$ and another at $\mathbf{x}_2$. Note that the number of the space-spin variables is reduced to one and two, respectively.

The total electron density $\rho_{\mathrm{N}}(\mathbf{r})$ and the spin density $\rho_{\mathrm{S}}(\mathbf{r})$ are expressed in terms of the spin-up and spin-down electron components, $\rho_{\kappa}(\mathbf{r})=\rho(\mathbf{r},\kappa)$, Eq.(2-43), $\kappa\in\{\uparrow,\downarrow\}\equiv\{\alpha,\beta\}$, as

$$\rho_{\mathrm{N}}(\mathbf{r})=\rho_{\alpha}(\mathbf{r})+\rho_{\beta}(\mathbf{r})=\sum_{\kappa}\rho(\mathbf{r},\kappa), \tag{2-87}$$

$$\rho_{\mathrm{S}}(\mathbf{r})=\rho_{\alpha}(\mathbf{r})-\rho_{\beta}(\mathbf{r}), \tag{2-88}$$

with their corresponding electron number and spin number

$$N=N_{\alpha}+N_{\beta}=\int\rho_{\mathrm{N}}(\mathbf{r})d\mathbf{r}=\int\rho_{\alpha}(\mathbf{r})d\mathbf{r}+\int\rho_{\beta}(\mathbf{r})d\mathbf{r}, \tag{2-89}$$

$$N_{\mathrm{S}}=N_{\alpha}-N_{\beta}=\int\rho_{\mathrm{S}}(\mathbf{r})d\mathbf{r}=\int\rho_{\alpha}(\mathbf{r})d\mathbf{r}-\int\rho_{\beta}(\mathbf{r})d\mathbf{r}\,. \tag{2-90}$$

The one- and two-electron density matrices obtained from the (Slater) determinantal N-electron wavefunction (see Eq.(2-20)) are especially simple. The determinantal 1-DM, $d_1=\gamma_1[\Phi]$, in terms of the orthonormal one-electron spin orbitals ϕ_i of Φ is:

$$d_1(\mathbf{x};\mathbf{x}')=d_1\left(\mathbf{x};\mathbf{x}';N,\{\phi_i\}\right)=\sum_{i}^{N}\phi_i(\mathbf{x})\phi_i^*(\mathbf{x}')\,. \tag{2-91}$$

The determinantal 2-DM, $d_2=\gamma_2[\Phi]$, is constructed using the d_1 as

$$d_2\left(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1',\mathbf{x}_2';[d_1]\right)=\frac{1}{2}\left\{d_1(\mathbf{x}_1;\mathbf{x}_1')d_1(\mathbf{x}_2;\mathbf{x}_2')-d_1(\mathbf{x}_1;\mathbf{x}_2')d_1(\mathbf{x}_1';\mathbf{x}_2)\right\}. \tag{2-92}$$

Note that $d_2\left[\gamma_1[\Psi]\right]=\gamma_2[\Psi]$ is true if and only if Ψ is a Slater determinant.

In general, the determinantal p-order reduced DM can be calculated from the 1-DM as a p-order determinant of d_1.[32] In the case of pure-state determinantal wavefunction, the spinor density is (see Eq.(2-91)

$$\rho(\mathbf{x})=d_1\left(\mathbf{x};\mathbf{x};N,\{\phi_i\}\right)=\sum_{i}^{N}\left|\phi_i(\mathbf{x})\right|^2. \tag{2-93}$$

II.D.2 Ensemble state

In the case when the knowledge of a system is incomplete, the expectation value of an observable $\hat{A}$ is expressed in terms of the ensemble DM $\hat{\Gamma}_N$ [44]

$$A=\left\langle\hat{A}\right\rangle_{\hat{\Gamma}_N}=\mathrm{tr}\,\hat{\Gamma}_N\hat{A} \tag{2-94}$$

(compare to Eq. (2-76) for the pure state).

In the N-electron Hilbert space, any ensemble density operator, $\hat{\Gamma}_N$, can be written as

$$\hat{\Gamma}_N = \sum_i g_i^N \left|\Psi_i\right\rangle\left\langle\Psi_i\right|, \; \left\langle\Psi_i\middle|\Psi_j\right\rangle = \delta_{ij}, \tag{2-95}$$

where pure states $\{|\Psi_i\rangle\}$ are elements of this space, while g_i^N is a probability of system to occur in the state $|\Psi_i\rangle$. The weights g_i^N satisfy

$$\mathrm{tr}\left(\hat{\Gamma}_K\right) = \sum_i g_i^N = 1; \qquad 0 \le g_i^N \le 1. \tag{2-96}$$

Using Eq.(2-95), the mean value, Eq.(2-94), is evaluated as

$$A = \sum_i g_i^N \left\langle\Psi_i\middle|\hat{A}\middle|\Psi_i\right\rangle. \tag{2-97}$$

While satisfying properties like Hermiticity, positive semi-definiteness and unit-trace one, $\hat{\Gamma}_N$, unlike $\hat{\gamma}_N$, does not possess the idempotency property

$$\begin{aligned}\hat{\Gamma}_N\hat{\Gamma}_N &= \sum_{ij} g_i^N g_j^N \left|\Psi_i\right\rangle\underbrace{\left\langle\Psi_i\middle|\Psi_j\right\rangle}_{\delta_{ij}}\left\langle\Psi_j\right| \\ &= \sum_i \left(g_i^N\right)^2 \left|\Psi_i\right\rangle\left\langle\Psi_i\right| \neq \sum_i g_i^N \left|\Psi_i\right\rangle\left\langle\Psi_i\right| = \hat{\Gamma}_N.\end{aligned} \tag{2-98}$$

The N-body ensemble DM in coordinate representation is

$$\begin{aligned}&\Gamma_N\left(\mathbf{x}_1,\cdots,\mathbf{x}_N;\mathbf{x}'_1,\cdots,\mathbf{x}'_N\right) \\ &\quad=\sum_i g_i^N \gamma_{N,i}\left(\mathbf{x}_1,\cdots,\mathbf{x}_N;\mathbf{x}'_1,\cdots,\mathbf{x}'_N\right) = \sum_i g_i^N \Psi_i\left(\mathbf{x}_1,\cdots,\mathbf{x}_N\right)\Psi_i^*\left(\mathbf{x}'_1,\cdots,\mathbf{x}'_N\right).\end{aligned} \tag{2-99}$$

Thus the ensemble DM is a weighted sum of the pure-state DMs.

The ensemble N-DM can be reduced into ensemble p-body DM

$$\Gamma_p^N\left(\mathbf{x}_1,\cdots,\mathbf{x}_p;\mathbf{x}'_1,\cdots,\mathbf{x}'_p\right) = \sum_i g_i^N \gamma_{p,i}^N\left(\mathbf{x}_1,\cdots,\mathbf{x}_p;\mathbf{x}'_1,\cdots,\mathbf{x}'_p\right) = \sum_i g_i^N \gamma_p^N\left[\Psi_i\right]. \tag{2-100}$$

In particular, Γ_2 and Γ_1 are given by

$$\Gamma_2^N(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1',\mathbf{x}_2') = \sum_i g_i^N \gamma_{2,i}^N(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1',\mathbf{x}_2'), \tag{2-101}$$

$$\Gamma_1^N(\mathbf{x}_1;\mathbf{x}_1') = \sum_i g_i^N \gamma_{1,i}^N(\mathbf{x}_1;\mathbf{x}_1'). \tag{2-102}$$

Eqs. (2-87)-(2-90) , (2-85) and (2-86) can be applied for the ensemble densities in the Hilbert space. Hence, the spinor ensemble density $\tilde{\rho}(\mathbf{x})$ is the diagonal of the ensemble 1-body reduced DM

$$\tilde{\rho}(\mathbf{x}_1) = \Gamma_1(\mathbf{x}_1;\mathbf{x}_1) = \tilde{\rho}(\mathbf{r}_1,\kappa_1;[\Gamma_N]), \tag{2-103}$$

while the spinless ensemble density $\tilde{\rho}_N(\mathbf{r})$ is

$$\tilde{\rho}_{\mathrm{N}}(\mathbf{r}) = \sum_\kappa \Gamma_1(\mathbf{r},\kappa;\mathbf{r},\kappa;[\Gamma_N]) = \sum_\kappa \tilde{\rho}_\kappa(\mathbf{r}), \tag{2-104}$$

where $\tilde{\rho}_\kappa(\mathbf{r}) = \tilde{\rho}(\mathbf{r},\kappa)$.

The definition of an ensemble DM operator can be extended from the Hilbert space to the Fock space (in general, the Fock space is defined as the direct sum of the Hilbert spaces, including the vacuum – the 0-electron space). An ensemble density operator $\hat{\Gamma}$ in the Fock space is defined as a mixture of Hilbert space contributions

$$\hat{\Gamma} = \sum_K \omega_K \hat{\Gamma}_K , \qquad K \in \{0,1,2,3,\cdots\} \tag{2-105}$$

with weights ω_K that satisfy

$$\sum_K \omega_K = 1; \qquad \omega_K \geq 0. \tag{2-106}$$

The average value of any operator in the Fock space, by generalization of Eq.(2-94), is

$$A = \langle \hat{A} \rangle_{\hat{\Gamma}} = \mathrm{Tr}\,\hat{\Gamma}\hat{A} = \sum_K \omega_K \langle \hat{A} \rangle_{\hat{\Gamma}_K} , \tag{2-107}$$

here "Tr" refers to trace in the Fock space. Particularly, for the electron number operator, $\hat{\mathcal{N}}$, the average electron number is

$$\mathcal{N} = \langle \hat{\mathcal{N}} \rangle_{\hat{\Gamma}} = \sum_K \omega_K K , \tag{2-108}$$

and can be any positive real number.

II.D.3 Illustrative examples

II.D.3.a General case for a molecule

The wavefunction which describes a molecular state is the solution of the Schrödinger equation with the Hamiltonian $\hat{H}[v]$ of the molecule. When the external electron-nuclei attraction potential $v(\mathbf{r})$ possesses some symmetry (is invariant with respect to a set of symmetry operations), the Hamiltonian commutes with the corresponding symmetry operators which are elements of the symmetry group G. Therefore the eigenstates of the Hamiltonian transform according to some irreducible representations g of G. Since the Hamiltonian commutes with the total-spin operators $\hat{\mathrm{S}}^2$ and $\hat{S}_{\mathrm{z}}$, the eigenstates of $\hat{H}$ are also eigenstates of these operators. Each pure state can be labeled: in the superscripts by the size $\mathcal{Z}$ of the Hilbert space $\mathcal{Y}^{\mathcal{Z}}$ to which it belongs and by the spin number, $\Sigma = 2S_{\mathrm{z}}$; in the subscripts by the ordinal number *i* of the Hamiltonian eigenstate in the $\mathcal{Z}$-particle Hilbert space and by the index *m* of the row of the irreducible representation to which the eigenstates belong, $|\Psi\rangle = |\Psi_{i,m}^{\mathcal{Z},\Sigma}\rangle$. The eigenstates are numbered so that $E_0^{\mathcal{Z}} < E_1^{\mathcal{Z}} < E_2^{\mathcal{Z}} < \ldots$.. The dependence on the external potential will be often suppressed.

The pure states can be grouped into a set of the eigenstates which belong to the $\mathcal{Z}$-particle Hilbert space

$$\mathcal{Y}^{\mathcal{Z}} = \left\{ \left| \Psi_{i,m}^{\mathcal{Z},\mathcal{S}} \right\rangle \;\middle|\; \hat{\mathcal{N}} \left| \Psi_{i,m}^{\mathcal{Z},\mathcal{S}} \right\rangle = \mathcal{Z} \left| \Psi_{i,m}^{\mathcal{Z},\mathcal{S}} \right\rangle \right\}. \tag{2-109}$$

This set can be split into subsets of $\mathcal{Y}^{\mathcal{Z}}$ consisting of states with the same i^{th} eigenvalue (multiplets)

$$\mathcal{Y}_i^{\mathcal{Z}} = \left\{ \left| \Psi_{i,m}^{\mathcal{Z},\mathcal{S}} \right\rangle \in \mathcal{Y}^{\mathcal{Z}} \;\middle|\; \hat{H} \left| \Psi_{i,m}^{\mathcal{Z},\mathcal{S}} \right\rangle = E_i^{\mathcal{Z}} \left| \Psi_{i,m}^{\mathcal{Z},\mathcal{S}} \right\rangle \right\}. \tag{2-110}$$

Note absence of degeneracy of the Hamiltonian eigenvalues with respect to eigenstates of $\hat{\mathrm{S}}^2$: $\hat{H}|\Psi_K\rangle = \hat{H}|\Psi_L\rangle \Rightarrow \hat{\mathrm{S}}^2|\Psi_K\rangle = \hat{\mathrm{S}}^2|\Psi_L\rangle$. This property is used to argue the omitting of the dependence $|\Psi_K\rangle$ on the total spin squared eigenvalues. Similarly, the belonging of two pure states to a given irreducible representation can be determined from the relation between their

Hamiltonian eigenfunctions: if $\hat{H}\left|\Psi_K\right\rangle = \hat{H}\left|\Psi_L\right\rangle$, then Ψ_K and Ψ_L transform according to the same irreducible representation.

The next level of splitting introduces subsets with the same spin number Σ

$$\mathcal{Y}_i^{\mathcal{Z},\Sigma} = \left\{ \left|\Psi_{i,m}^{\mathcal{Z},\Sigma}\right\rangle \;\middle|\; \left|\Psi_{i,m}^{\mathcal{Z},\Sigma}\right\rangle \in \mathcal{Y}_i^{\mathcal{Z}}, 2\hat{S}_z \left|\Psi_{i,m}^{\mathcal{Z},\Sigma}\right\rangle = \Sigma \left|\Psi_{i,m}^{\mathcal{Z},\Sigma}\right\rangle \right\} \qquad (2\text{-}111)$$

or subsets with the *k*th row of the irreducible representation g of the group G

$$\mathcal{Y}_{i,k}^{\mathcal{Z}} = \left\{ \left|\Psi_{i,k}^{\mathcal{Z},\mathcal{S}}\right\rangle \;\middle|\; \left|\Psi_{i,k}^{\mathcal{Z},\mathcal{S}}\right\rangle \in \mathcal{Y}_i^{\mathcal{Z}}, \hat{R}_{mm}^{G,g} \left|\Psi_{i,k}^{\mathcal{Z},\mathcal{S}}\right\rangle = \delta_{mk} \left|\Psi_{i,k}^{\mathcal{Z},\mathcal{S}}\right\rangle \right\}. \qquad (2\text{-}112)$$

Here $\hat{R}_{mm}^{G,g}$ is the corresponding projection operator.

There are disjoining relations between sets

$$\begin{array}{lll} \mathcal{Y}^{\mathcal{Z}} \cap \mathcal{Y}^{\mathcal{Z}'} = \varnothing & \text{for} & \mathcal{Z} \neq \mathcal{Z}', \\ \mathcal{Y}_i^{\mathcal{Z}} \cap \mathcal{Y}_j^{\mathcal{Z}} = \varnothing & \text{for} & j \neq i, \\ \mathcal{Y}_i^{\mathcal{Z},\Sigma} \cap \mathcal{Y}_i^{\mathcal{Z},\Sigma'} = \varnothing & \text{for} & \Sigma \neq \Sigma', \\ \mathcal{Y}_{i,k}^{\mathcal{Z}} \cap \mathcal{Y}_{i,l}^{\mathcal{Z}} = \varnothing & \text{for} & l \neq k, \end{array} , \qquad (2\text{-}113)$$

and the intersection of $\mathcal{Y}_i^{\mathcal{Z},\Sigma}$ and $\mathcal{Y}_{i,k}^{\mathcal{Z}}$,

$$\mathcal{Y}_i^{\mathcal{Z},\Sigma} \cap \mathcal{Y}_{i,k}^{\mathcal{Z}} = \left|\Psi_{i,k}^{\mathcal{Z},\Sigma}\right\rangle \qquad (2\text{-}114)$$

is a single-element set – a pure state.

Based on relations (2-113) and (2-114), and the union relations

$$\mathcal{Y}_i^{\mathcal{Z}} = \bigcup\nolimits_k \mathcal{Y}_{i,k}^{\mathcal{Z}} = \bigcup\nolimits_\Sigma \mathcal{Y}_i^{\mathcal{Z},\Sigma}, \qquad (2\text{-}115)$$

the degeneracy of the given i^{th} eigenvalues of $\hat{H}$ is

$$d_i^{\mathcal{Z}} = \#\mathcal{Y}_i^{\mathcal{Z}} = d_i^{\mathcal{Z},\Sigma} d_{i,k}^{\mathcal{Z}} = d_g^{\mathcal{Z},i} d_{\mathrm{S}}^{\mathcal{Z},i}, \qquad (2\text{-}116)$$

where $\#\mathcal{Y}$ means the cardinality of the set $\mathcal{Y}$ – the number of elements in the set, $d_i^{\mathcal{Z},\Sigma}$ and $d_{i,k}^{\mathcal{Z}}$ are the degeneracies of the i^{th} eigenvalue in the configuration and spin subspaces:

$$d_i^{\mathcal{Z},\Sigma} = \#\mathcal{Y}_i^{\mathcal{Z},\Sigma} = d_g^{\mathcal{Z},i}, \qquad (2\text{-}117)$$

$$d_{i,k}^{\mathcal{Z}} = \#\mathcal{Y}_{i,k}^{\mathcal{Z}} = d_{\mathrm{S}}^{\mathcal{Z},i} = 2S_{\mathrm{T}}^{\mathcal{Z},i} + 1, \qquad (2\text{-}118)$$

where $S_{\mathrm{T}}^{\mathcal{Z},i}$ is the total spin number of a pure state $\left|\Psi_{i,m}^{\mathcal{Z},\varSigma}\right\rangle$, i.e. satisfying

$$\hat{\mathrm{S}}^2\left|\Psi_{i,m}^{\mathcal{Z},\varSigma}\right\rangle = S_{\mathrm{T}}^{\mathcal{Z},i}\left(S_{\mathrm{T}}^{\mathcal{Z},i}+1\right)\left|\Psi_{i,m}^{\mathcal{Z},\varSigma}\right\rangle, \qquad (2\text{-}119)$$

and $d_g^{\mathcal{Z},i}$ is the dimension of the irreducible representation g describing the spatial symmetry of eigenstates corresponding to the eigenvalue $E_i^{\mathcal{Z}}$.

Using Eqs.(2-79) and (2-85), the density of the molecular pure state is

$$\rho_{i,k}^{\mathcal{Z},\varSigma}(\mathbf{x}) = \gamma_1\left(\mathbf{x};\mathbf{x};\left[\Psi_{i,k}^{\mathcal{Z},\varSigma}\right]\right). \qquad (2\text{-}120)$$

The equi-ensemble density operator constructed from the $\mathcal{Y}_{i,k}^{\mathcal{Z}}$ states is

$$\hat{\varGamma}_{i,k}^{\mathcal{Z}} = \frac{1}{d_{\mathrm{S}}^{\mathcal{Z},i}}\sum_{\varSigma=-2S_{\mathrm{T}}^{\mathcal{Z},i}}^{2S_{\mathrm{T}}^{\mathcal{Z},i},2}\left|\Psi_{i,k}^{\mathcal{Z},\varSigma}\right\rangle\left\langle\Psi_{i,k}^{\mathcal{Z},\varSigma}\right|, \qquad (2\text{-}121)$$

and the corresponding ensemble density, Eq.(2-103), is

$$\tilde{\rho}_{i,k}^{\mathcal{Z}}(\mathbf{x}) = \gamma_1\left(\mathbf{x};\mathbf{x};\left[\varGamma_{i,k}^{\mathcal{Z}}\right]\right). \qquad (2\text{-}122)$$

In a similar way, the equi-ensemble density operator constructed from the $\mathcal{Y}_i^{\mathcal{Z},\varSigma}$ states is

$$\hat{\varGamma}_i^{\mathcal{Z},\varSigma} = \frac{1}{d_g^{\mathcal{Z},i}}\sum_m^{d_g^{\mathcal{Z},i}}\left|\Psi_{i,m}^{\mathcal{Z},\varSigma}\right\rangle\left\langle\Psi_{i,m}^{\mathcal{Z},\varSigma}\right|, \qquad (2\text{-}123)$$

and the ensemble density, Eq.(2-103), is

$$\tilde{\rho}_i^{\mathcal{Z},\varSigma}(\mathbf{x}) = \gamma_1\left(\mathbf{x};\mathbf{x};\left[\varGamma_i^{\mathcal{Z},\varSigma}\right]\right). \qquad (2\text{-}124)$$

Finally, the equi-ensemble density operator constructed from the $\mathcal{Y}_i^{\mathcal{Z}}$ states is

$$\hat{\varGamma}_i^{\mathcal{Z}} = \frac{1}{d_i^{\mathcal{Z}}}\sum_{\varSigma=-2S_{\mathrm{T}}^{\mathcal{Z},i}}^{2S_{\mathrm{T}}^{\mathcal{Z},i},2}\sum_m^{d_g^{\mathcal{Z},i}}\left|\Psi_{i,m}^{\mathcal{Z},\varSigma}\right\rangle\left\langle\Psi_{i,m}^{\mathcal{Z},\varSigma}\right| = \frac{1}{d_S^{\mathcal{Z},i}}\sum_{\varSigma=-2S_{\mathrm{T}}^{\mathcal{Z},i}}^{2S_{\mathrm{T}}^{\mathcal{Z},i},2}\hat{\varGamma}_i^{\mathcal{Z},\varSigma} = \frac{1}{d_g^{\mathcal{Z},i}}\sum_m^{d_g^{\mathcal{Z},i}}\hat{\varGamma}_{i,m}^{\mathcal{Z}}. \qquad (2\text{-}125)$$

The corresponding multiplet density is

$$\tilde{\rho}_i^{\mathcal{Z}}(\mathbf{x}) = \gamma_1\left(\mathbf{x};\mathbf{x};\left[\varGamma_i^{\mathcal{Z}}\right]\right). \qquad (2\text{-}126)$$

Construction of the ensemble density for molecule is summarized in Fig. II-1. The spinless and spin density can be obtained according to Eq.(2-87) and Eq.(2-88), respectively.

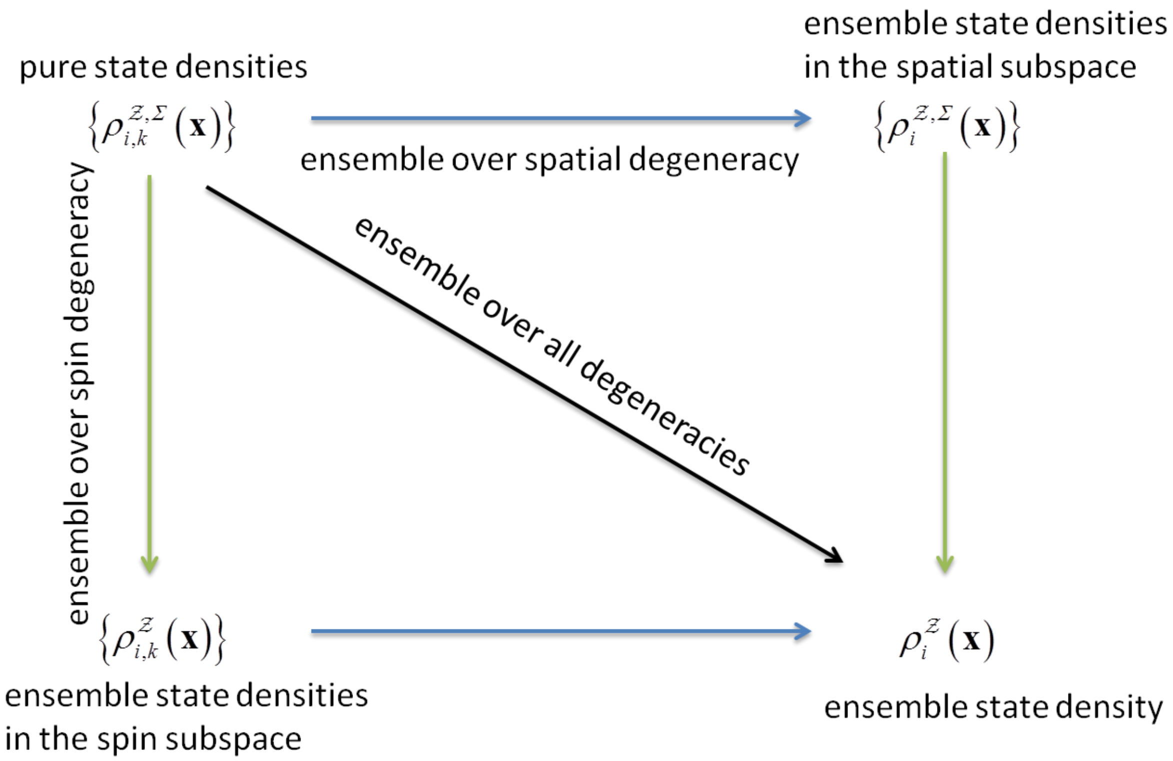
pure state densities
$\{\rho^{\mathcal{Z},\Sigma}_{i,k}(\mathbf{x})\}$
ensemble state densities
in the spatial subspace
$\{\rho^{\mathcal{Z},\Sigma}_{i}(\mathbf{x})\}$
ensemble over spatial degeneracy
ensemble over all degeneracies
ensemble over spin degeneracy
$\{\rho^{\mathcal{Z}}_{i,k}(\mathbf{x})\}$
ensemble state densities
in the spin subspace
$\rho^{\mathcal{Z}}_{i}(\mathbf{x})$
ensemble state density

Fig. II-1. Ensemble density construction.

II.D.3.b General case for atoms

In the atomic case, any pure state can be unambiguously labeled by the number of electrons $\mathcal{Z}$ (equal the atomic number for neutral state), by the ordinal number i of eigenvalues of the Hamiltonian, by the eigenvalue of the z-component $\hat{L}_z$ of the total orbital angular momentum operator $\hat{\mathrm{L}}$, and by Σ — the doubled eigenvalues of the z-component $\hat{S}_z$ of the spin operator $\hat{\mathrm{S}}$. From Eq.(2-120), the pure state density is

$$\rho^{\mathcal{Z},\Sigma,L_z,i}(\mathbf{x}) = \gamma_1\left(\mathbf{x};\mathbf{x};\left[\Psi^{\mathcal{Z},\Sigma}_{i,L_z}\right]\right), \qquad (2\text{-}127)$$

here the index of the row of the irreducible representation g of a molecule is replaced by the eigenvalue of the $\hat{L}_z$ operator. As previously discussed, absence of degeneracy of the Hamiltonian eigenvalues with respect to $\hat{\mathrm{L}}^2$ is used to argue omitting the dependence of the state on $\hat{L}^2$ eigenvalues. Following Eq.(2-104), the spinless density is

$$\rho_{\mathrm{N}}^{\mathcal{Z},\Sigma,L_z,i}\left(\mathbf{r}\right)=\rho_{\alpha}^{\mathcal{Z},\Sigma,L_z,i}\left(\mathbf{r}\right)+\rho_{\beta}^{\mathcal{Z},\Sigma,L_z,i}\left(\mathbf{r}\right) \tag{2-128}$$

and according to Eq.(2-88), the spin density is

$$\rho_{\mathrm{S}}^{\mathcal{Z},\Sigma,L_z,i}\left(\mathbf{r}\right)=\rho_{\alpha}^{\mathcal{Z},\Sigma,L_z,i}\left(\mathbf{r}\right)-\rho_{\beta}^{\mathcal{Z},\Sigma,L_z,i}\left(\mathbf{r}\right). \tag{2-129}$$

The ensemble densities can be constructed using Eqs. (2-121)-(2-126). However, in practical application usually only the solution for the highest component of the spin multiplet, $\Sigma=2S_{\mathrm{T}}^{\mathcal{Z},i}$, and for $L_z=0$ component of the angular momentum multiplet (with the *z*-axis assumed to be a principal symmetry axis) is at our disposal.

Basing on the spin structure of 1-DM,[44] the densities of the remaining spin-multiplet components can be constructed in terms of the spinless density and the spin density of this highest component of the spin multiplet:

$$\begin{aligned}&\left(\rho_{\alpha}^{\mathcal{Z},\Sigma,L_z,i}\left(\mathbf{r}\right),\rho_{\beta}^{\mathcal{Z},\Sigma,L_z,i}\left(\mathbf{r}\right)\right)\\&=\frac{1}{2}\begin{pmatrix}\rho_{\mathrm{N}}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},L_z,i}\left(\mathbf{r}\right)\\+\dfrac{\Sigma}{2S_{\mathrm{T}}^{\mathcal{Z},i}}\rho_{\mathrm{S}}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},L_z,i}\left(\mathbf{r}\right),\rho_{\mathrm{N}}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},L_z,i}\left(\mathbf{r}\right)-\dfrac{\Sigma}{2S_{\mathrm{T}}^{\mathcal{Z},i}}\rho_{\mathrm{S}}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},L_z,i}\left(\mathbf{r}\right)\end{pmatrix}.\end{aligned} \tag{2-130}$$

From Eqs.(2-121) and (2-122), using Eq.(2-130), the spin-component densities of the ensemble over the spin degeneracy are

$$\begin{aligned}\rho_{\alpha/\beta}^{\mathcal{Z},L_z,i}\left(\mathbf{r}\right)&=\frac{1}{2}\frac{1}{\left(2S_{\mathrm{T}}^{\mathcal{Z},i}+1\right)}\sum_{\Sigma=-2S_{\mathrm{T}}^{\mathcal{Z},i}}^{2S_{\mathrm{T}}^{\mathcal{Z},i},2}\left(\rho_{\mathrm{N}}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},L_z,i}\left(\mathbf{r}\right)\pm\frac{\Sigma}{2S_{\mathrm{T}}^{\mathcal{Z},i}}\rho_{\mathrm{S}}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},L_z,i}\left(\mathbf{r}\right)\right)\\&=\frac{1}{2}\rho_{\mathrm{N}}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},L_z,i}\left(\mathbf{r}\right),\end{aligned} \tag{2-131}$$

happen to be independent of α and β.

This result yields

$$\rho_{\mathrm{N}}^{\mathcal{Z},L_z,i}\left(\mathbf{r}\right)=\left(\rho_{\alpha}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},L_z,i}\left(\mathbf{r}\right)+\rho_{\beta}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},L_z,i}\left(\mathbf{r}\right)\right), \tag{2-132}$$

$$\rho_{\mathrm{S}}^{\mathcal{Z},L_z,i}\left(\mathbf{r}\right)=0. \tag{2-133}$$

In analogy with Eq.(2-119) for a spin-space property, we have a configuration-space property for atoms

$$\hat{\mathrm{L}}^2\left|\Psi_{i,L_z}^{\mathcal{Z},\Sigma}\right\rangle=L_{\mathrm{T}}\left(L_{\mathrm{T}}+1\right)\left|\Psi_{i,L_z}^{\mathcal{Z},\Sigma}\right\rangle, \tag{2-134}$$

So the degeneracy in this space is $d = 2L_{\mathrm{T}} + 1$. The density of equi-ensemble in this space is the average of $2L_{\mathrm{T}} + 1$ pure-state L_{z}-labeled densities, each one proportional to $\left|Y_{L_{\mathrm{T}},L_{\mathrm{z}}}\right|^2$ (in terms of the spherical harmonics). It is convenient to transform the set $\left\{Y_{L_{\mathrm{T}},L_{\mathrm{z}}}\right\}$ into another orthonormal set of real functions. In the case of $L_{\mathrm{T}} = 1$ (the P-type angular momentum), the $L_{\mathrm{z}} = 0$ function is among transformed ones, other transformed functions are specific rotations of this one, so

$$\begin{aligned}\rho_{\kappa}^{\mathcal{Z},\Sigma,i}\left(\mathbf{r}\right) &= \frac{1}{3}\left(\rho_{\kappa}^{\mathcal{Z},\Sigma,L_{\mathrm{z}}=-1,i}\left(\mathbf{r}\right) + \rho_{\kappa}^{\mathcal{Z},\Sigma,L_{\mathrm{z}}=0,i}\left(\mathbf{r}\right) + \rho_{\kappa}^{\mathcal{Z},\Sigma,L_{\mathrm{z}}=+1,i}\left(\mathbf{r}\right)\right) \\ &= \frac{1}{3}\left(\rho_{\kappa}^{\mathcal{Z},\Sigma,L_{\mathrm{z}}=0,i}\left(x,y,z\right) + \rho_{\kappa}^{\mathcal{Z},\Sigma,L_{\mathrm{z}}=0,i}\left(y,z,x\right) + \rho_{\kappa}^{\mathcal{Z},\Sigma,L_{\mathrm{z}}=0,i}\left(z,x,y,\right)\right).\end{aligned} \tag{2-135}$$

Note that $\rho_{\kappa}^{\mathcal{Z},\Sigma,i}$ shows the spherical symmetry, while pure-state densities are lacking it. The multiplet density can be constructed from $\rho_{\kappa}^{\mathcal{Z},\Sigma,i}$ of $\Sigma = 2S_{\mathrm{T}}^{\mathcal{Z},i}$ using Eq.(2-131): the spin components are

$$\rho_{\alpha}^{\mathcal{Z},i}\left(\mathbf{r}\right) = \rho_{\beta}^{\mathcal{Z},i}\left(\mathbf{r}\right) = \frac{1}{2}\rho_{\mathrm{N}}^{\mathcal{Z},S_{\mathrm{T}}^{\mathcal{Z},i},i}\left(\mathbf{r}\right), \tag{2-136}$$

therefore the spinless density and spin density are

$$\rho_{\mathrm{N}}^{\mathcal{Z},i}\left(\mathbf{r}\right) = \rho_{\mathrm{N}}^{\mathcal{Z},2S_{\mathrm{T}}^{\mathcal{Z},i},i}\left(\mathbf{r}\right), \tag{2-137}$$

$$\rho_{\mathrm{S}}^{\mathcal{Z},i}\left(\mathbf{r}\right) = 0. \tag{2-138}$$

Construction of the ensemble density will be illustrated on examples of boron, carbon and nitrogen atoms. Their total orbital angular momentum value is $L_{\mathrm{T}} = 1$ for B and C, $L_{\mathrm{T}} = 0$ for N.

Boron with $\mathcal{N} = 5$ has an electronic configuration $1s^2 2s^2 2p^1$. It is doublet in the spin space, $S_{\mathrm{T}}^{5,0} = 1/2$. From ordinary computational calculation, only the ground state density is available, $\rho^{5,1,0,0}\left(\mathbf{x}\right)$, for $\Sigma = 2S_{\mathrm{T}}^{5,0} = 1$, $L_{\mathrm{z}} = 0$, and $i = 0$. Using Eqs. (2-127)-(2-129), the spin components, the spinless density and the spin density are calculable. The doublet complementary is accessible from Eq.(2-130)

$$\begin{aligned}\left(\rho_{\alpha}^{5,-1,0,0}\left(\mathbf{r}\right), \rho_{\beta}^{5,-1,0,0}\left(\mathbf{r}\right)\right) &= \frac{1}{2}\left(\rho_{\mathrm{N}}^{5,1,0,0}\left(\mathbf{r}\right) - \rho_{\mathrm{S}}^{5,1,0,0}\left(\mathbf{r}\right), \rho_{\mathrm{N}}^{5,1,0,0}\left(\mathbf{r}\right) + \rho_{\mathrm{S}}^{5,1,0,0}\left(\mathbf{r}\right)\right) \\ &= \left(\rho_{\beta}^{5,1,0,0}\left(\mathbf{r}\right), \rho_{\alpha}^{5,1,0,0}\left(\mathbf{r}\right)\right),\end{aligned} \tag{2-139}$$

and this yields

$$\rho_{\mathrm{N}}^{5,-1,0,0}(\mathbf{r}) = \rho_{\mathrm{N}}^{5,1,0,0}(\mathbf{r}), \qquad \rho_{\mathrm{S}}^{5,-1,0,0}(\mathbf{r}) = -\rho_{\mathrm{S}}^{5,1,0,0}(\mathbf{r}). \tag{2-140}$$

The equi-ensemble in the spin space has the same spinless density as the spinless density for the spin-components but the spin density is equal zero (from Eq.(2-133)). The equi-ensemble density in the configurational space is calculated using Eq.(2-135)

$$\rho_{\kappa}^{5,\Sigma=1,0}(\mathbf{r}) = \frac{1}{3}\left(\rho_{\kappa}^{5,1,0,0}(x,y,z) + \rho_{\kappa}^{5,1,0,0}(y,z,x) + \rho_{\kappa}^{5,1,0,0}(z,x,y)\right). \tag{2-141}$$

The multiplet spinless density is

$$\rho_{\mathrm{N}}^{5,0}(\mathbf{r}) = \frac{1}{3}\left(\rho_{\mathrm{N}}^{5,1,0,0}(x,y,z) + \rho_{\mathrm{N}}^{5,1,0,0}(y,z,x) + \rho_{\mathrm{N}}^{5,1,0,0}(z,x,y)\right), \tag{2-142}$$

and the multiplet spin density is zero in the whole space.

Carbon with $\mathcal{N} = 6$ has an electronic configuration $1s^2 2s^2 2p^2$. Its ground state $(i = 0)$ energy is 9-fold degenerate because $S_{\mathrm{T}}^{6,0} = L_{\mathrm{T}} = 1$. The pure state densities and the ensemble densities are calculable in the same way as in the boron case. We have following relations

$$\rho_{\beta/\alpha}^{6,\Sigma=-2,L_Z=0,i}(\mathbf{r}) = \rho_{\alpha/\beta}^{6,\Sigma=2,L_Z=0,i}(\mathbf{r});\ \rho_{\alpha/\beta}^{6,\Sigma=0,L_Z=0,i}(\mathbf{r}) = \frac{1}{2}\rho_{\mathrm{N}}^{6,\Sigma=2,L_Z=0,i}(\mathbf{r}), \tag{2-143}$$

$$\begin{aligned}\rho_{\mathrm{N}}^{6,\Sigma=2,L_Z=0,0}(\mathbf{r}) &= \rho_{\mathrm{N}}^{6,\Sigma=0,L_Z=0,0}(\mathbf{r}) = \rho_{\mathrm{N}}^{6,\Sigma=-2,L_Z=0,0}(\mathbf{r}) \\ &= \left(\rho_{\alpha}^{6,\Sigma=2,L_Z=0,0}(\mathbf{r}) + \rho_{\beta}^{6,\Sigma=2,L_Z=0,0}(\mathbf{r})\right),\end{aligned} \tag{2-144}$$

$$\begin{aligned}&\rho_{\mathrm{S}}^{6,\Sigma=-2,L_Z=0,0}(\mathbf{r}) = -\rho_{\mathrm{S}}^{6,\Sigma=2,L_Z=0,0}(\mathbf{r}) = -\left(\rho_{\alpha}^{6,\Sigma=2,L_Z=0,0} - \rho_{\beta}^{6,\Sigma=2,L_Z=0,0}\right); \\ &\rho_{\mathrm{S}}^{6,\Sigma=0,L_Z=0,0}(\mathbf{r}) = 0.\end{aligned} \tag{2-145}$$

As in the boron case, the multiplet spinless density is simply the average of the sum of $\rho_{\mathrm{N}}^{6,\Sigma=2,L_Z=0,0}$ and its rotations in the configuration space

$$\rho_{\mathrm{N}}^{6,0}(\mathbf{r}) = \frac{1}{3}\left(\rho_{\mathrm{N}}^{6,2,0,0}(x,y,z) + \rho_{\mathrm{N}}^{6,2,0,0}(y,z,x) + \rho_{\mathrm{N}}^{6,2,0,0}(z,x,y)\right), \tag{2-146}$$

$\rho_{\mathrm{S}}^{6,0}(\mathbf{r})$ is simply zero.

In nitrogen case, the ground state is 4-fold degenerate, $S_{\mathrm{T}}^{\mathcal{Z},0} = 3/2$, $L_{\mathrm{T}} = 0$. There is no degeneracy in the configuration space. The pure state densities and the ensemble densities are calculable in as in preceding cases. Note that the pure state density for any $\Sigma \in \{-3,-1,+1,+3\}$ is not simply spinless density of $\Sigma = +3$, but Eq.(2-130) yields

$$\begin{aligned}&\left(\rho_{\alpha}^{7,\Sigma,L_Z=0,0}(\mathbf{r}),\rho_{\beta}^{7,\Sigma,L_Z=0,0}(\mathbf{r})\right)\\&=\frac{1}{2}\left(\rho_{\mathrm{N}}^{7,3,0,0}(\mathbf{r})+\frac{\Sigma}{3}\rho_{\mathrm{S}}^{7,3,0,0}(\mathbf{r}),\rho_{\mathrm{N}}^{7,3,0,0}(\mathbf{r})-\frac{\Sigma}{3}\rho_{\mathrm{S}}^{7,3,0,0}(\mathbf{r})\right),\end{aligned} \tag{2-147}$$

e.g. for $\Sigma = +1$, we have

$$\begin{aligned}&\left(\rho_{\alpha}^{7,\Sigma=1,L_Z=0,0}(\mathbf{r}),\rho_{\beta}^{7,\Sigma=1,L_Z=0,0}(\mathbf{r})\right)\\&=\left(\frac{2}{3}\rho_{\alpha}^{7,3,0,0}(\mathbf{r})+\frac{1}{3}\rho_{\beta}^{7,3,0,0}(\mathbf{r}),\frac{1}{3}\rho_{\alpha}^{7,3,0,0}(\mathbf{r})+\frac{2}{3}\rho_{\beta}^{7,3,0,0}(\mathbf{r})\right).\end{aligned} \tag{2-148}$$

In this case the multiplet spinless density is very simple

$$\rho_{\mathrm{N}}^{7,0}(\mathbf{r})=\rho_{\mathrm{N}}^{7,3,0,0}(\mathbf{r}), \tag{2-149}$$

while the spin density vanishes, $\rho_{\mathrm{S}}^{7,0}(\mathbf{r})=0$.

The ensemble densities in the Fock space are constructed in a similar way.

Chapter III. Chemical Space and Reactivity Indices

The development of methods for systematic and rational classification of chemical compounds and the search for compounds with desired properties is a fundamental task in chemical, material and pharmaceutical research.[17-19,45-47] Advances in this field led to idea of *chemical space* (CS) which was defined as set of all possible combinations of chemical elements building all geometric isomers,[20,48] but the purposeful design of molecules with optimized properties is daunting because the number of accessible stable molecules is immense.[49,50] From the quantum chemical point of view, exploration of the complete CS is a challenging task as computational timings become prohibitively large (assuming that to predict some property of a molecule at least a single-point calculation has to be performed, often demanding high level ab-initio methods). Thus, the straightforward exploration of CS becomes a nearly impossible procedure and more sophisticated methods like simulated annealing,[19] genetic algorithms,[51] linear combinations of atomic potentials,[45] or optimization based on alchemical gradients,[35-37,52] have to be used to cope with this problem in an efficient way.

A general description of chemical space involving changes in composition and geometry is possible using density functional reactivity theory (DFRT), often called conceptual density functional theory (DFT).[38-40,53-55] Conceptual DFT is based on the study of the $E = E\left[N, v\right]$ (the energy vs. number of electrons (N) and external potential $v(\mathbf{r})$ functional), more precisely on the change the functional undergoes when the number of electrons and /or the external potential is varied, concretized by the corresponding response functions. Traditionally DFRT has been concerned primarily with the prediction of phenomena associated with electron transfer, i.e. responses to changes in the number of electrons. Recently, numerous workers have begun to focus on the response of a system to a change in the external potential,[56-58] which for example results from shifts of the nuclear positions in a molecule, from the approach of reagents or from changing the nuclear charges. Techniques have been developed for the numerical and analytical evaluation of the corresponding response functions including the so called linear response function which recently received a lot of attention by some authors.[59-65] Changing the location of the point charges can be useful in geometry optimizations.[66] Addition of a positive point charge on the periphery of the molecule[56,66] is useful for discerning its nucleophilicity and its Brønsted-Lowry acidity.

In the present chapter, the concept of the chemical space will be introduced at first. After some general ideas of the conceptual DFT will be briefly recalled. Finally, the concept of the alchemical derivative will be presented.

III.A Chemical space

The specification of the relative positions of all atoms in *M*-dimensional space, is termed as *M-dimensional conformational space,* is defined as

$$\mathcal{C}_M \equiv \left\{ \left(\mathbf{R}^M, \mathbf{Z}^M\right) \middle| \mathbf{R}^M \in \mathcal{R}_M, \mathbf{Z}^M \in \mathcal{Z}_M \right\} \tag{3-1}$$

with $\mathcal{R}_M$ – a 3*M-dimensional spatial configuration space*

$$\mathcal{R}_M \equiv \left\{ \mathbf{R}^M = \left(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_M\right) \middle| \forall A, B \in \{1,2,..,M\}, A \neq B : \mathbf{R}_A \neq \mathbf{R}_B \right\}, \tag{3-2}$$

and $\mathcal{Z}_M$ – a *M-dimensional charge configuration space*

$$\mathcal{Z}_M \equiv \left\{ \mathbf{Z}^M = \left(Z_1, Z_2, ..., Z_M\right) \middle| Z_A \in \mathbb{R}, Z_A \geq 0 \right\}, \tag{3-3}$$

where $\mathbf{R}_A$ and Z_A denote the nuclear locations and the corresponding nuclear charges, respectively. A fractional charge is allowed in the present description. The site with the charge equal zero is called a vacancy and collapsed nuclear coordinates are excluded from $\mathcal{R}_M$.

To each point of the conformational space $\left(\mathbf{R}^M, \mathbf{Z}^M\right)$ corresponds an external potential, defined as

$$v\left(\mathbf{r}, \mathbf{Z}^M, \mathbf{R}^M\right) = -\sum_A \frac{Z_A}{\left|\mathbf{r} - \mathbf{R}_A\right|} \ , \tag{3-4}$$

The nuclear-nuclear repulsion energy is

$$V_{\mathrm{nn}}\left[\mathbf{Z}^M, \mathbf{R}^M\right] = \sum_A \sum_{B>A} \frac{Z_A Z_B}{\left|\mathbf{R}_A - \mathbf{R}_B\right|}, \tag{3-5}$$

and the total proton number is defined as: $N_p\left[\mathbf{Z}^M\right] = \sum_A Z_A$.

Every point of the conformational space can be associated with some electronic configuration characterized by the electron number *N* and some characteristics in the spin space, e.g. singlet, doublet, etc (see Fig. III-1). This definition can be extended to the systems with noninteger electron number using the ensemble approach.[67-69]

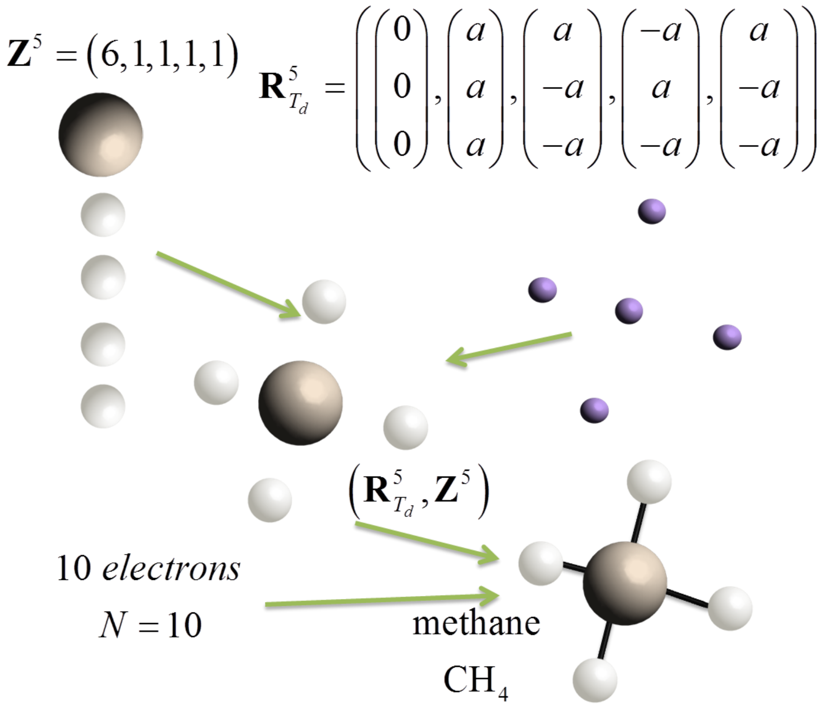
$\mathbf{Z}^5 = (6,1,1,1,1)$
$\mathbf{R}^5_{T_d} = \left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}a\\a\\a\end{pmatrix}, \begin{pmatrix}a\\-a\\-a\end{pmatrix}, \begin{pmatrix}-a\\a\\-a\end{pmatrix}, \begin{pmatrix}a\\-a\\-a\end{pmatrix} \right)$
$\left(\mathbf{R}^5_{T_d}, \mathbf{Z}^5\right)$
10 *electrons*
$N = 10$
methane
CH_4

Fig. III-1 Methane in the 5-dimensional conformational space

The total energy W of a system, the sum of the electronic energy E and the nuclear-nuclear repulsion energy V_{nn} (without another external filed) is

$$W\left[N, \mathbf{Z}^M, \mathbf{R}^M\right] \equiv E\left[N, v\left[\mathbf{Z}^M, \mathbf{R}^M\right]\right] + V_{\mathrm{nn}}\left[\mathbf{Z}^M, \mathbf{R}^M\right]. \tag{3-6}$$

Consequently, the independent variables are the coordinates in the conformational space $\left(\mathbf{R}^M, \mathbf{Z}^M\right)$ and the electron number, N.

Moving from one point to another in the chemical space is characterized by a change of the position vector

$$d\mathbf{R}^M = \left(d\mathbf{R}_1, d\mathbf{R}_2, \ldots, d\mathbf{R}_M\right), \tag{3-7}$$

the change of the charge vector, the so-called *transmutation vector*

$$d\mathbf{Z}^M = \left(dZ_1, dZ_2, \ldots, dZ_M\right), \tag{3-8}$$

and the change in the electron number dN. So “the journey” in the chemical space can be expressed as the change vector $\left(d\mathbf{R}^M, d\mathbf{Z}^M, dN\right)$. The change in the total energy in the canonical ensemble, can be computed using the functional Taylor series as

$$dW = W\left[N + dN, \mathbf{Z}^M + d\mathbf{Z}^M, \mathbf{R}^M + d\mathbf{R}^M\right] - W\left[N, \mathbf{Z}^M, \mathbf{R}^M\right] \tag{3-9}$$

$$= \left(\frac{\partial W}{\partial \mathbf{R}^M}\right) d\mathbf{R}^M + \frac{1}{2}\left(d\mathbf{R}^M\right)^{\mathrm{T}} \left(\frac{\partial^2 W}{\partial \mathbf{R}^M \partial \mathbf{R}^M}\right) d\mathbf{R}^M \tag{3-10}$$

$$+\left(\frac{\partial W}{\partial \mathbf{Z}^M}\right)d\mathbf{Z}^M+\frac{1}{2}\left(d\mathbf{Z}^M\right)^{\mathrm{T}}\left(\frac{\partial^2 W}{\partial \mathbf{Z}^M\partial \mathbf{Z}^M}\right)d\mathbf{Z}^M \tag{3-11}$$

$$+\left(\frac{\partial W}{\partial N}\right)dN+\frac{1}{2}\left(\frac{\partial^2 W}{\partial N^2}\right)dN^2 \tag{3-12}$$

$$+\left(\frac{\partial^2 W}{\partial \mathbf{R}^M\partial N}\right)d\mathbf{R}^M dN+\left(\frac{\partial^2 W}{\partial \mathbf{Z}^M\partial N}\right)d\mathbf{Z}^M dN+\left(d\mathbf{Z}^M\right)^{\mathrm{T}}\left(\frac{\partial^2 W}{\partial \mathbf{R}^M\partial \mathbf{R}^M}\right)d\mathbf{R}^M . \tag{3-13}$$

This series is truncated after second-order terms. The derivatives are the response functions to perturbations in these variables. Many of these functions have been linked to chemical concepts, (previously) often vaguely defined but readily used by chemists. The derivatives of the total energy with respect to the displacements of the atoms, Eq.(3-10) are well know and used, e.g. for the geometry optimization. The derivative with respect to the nuclear charge will be deeply discussed in this chapter.

Traditionally DFRT has been concerned primarily with the prediction of phenomena associated with electron transfer, i.e. responses to changes in the number of electrons, Eq.(3-12). These derivatives and other local reactivity indices will be shortly recalled in the next section. From the mixed derivatives, Eq.(3-13) only derivatives with respect to the electron number and nucleus displacements were studied. These are so called nuclear derivative, e.g. nuclear Fukui function[70-72] and nuclear stiffness.[73]

III.B The DFT-based chemical reactivity indices

The indices from conceptual DFT can be introduced in a straightforward way by considering the Taylor-like expansion for a system's energy. In this series expansion, the energy of the isolated system is the zeroth order term and the higher order terms correspond to perturbations in the energy either due to a change in the number of electrons or a change in the external potential. The derivatives appearing in the higher order terms reflect the initial energy change occurring in a chemical reaction and they have been identified as the reactivity indices from conceptual DFT. One of the goals of conceptual DFT is to describe and understand an atom's or a molecule's chemical behavior in terms of these indicators. The indices appearing in the perturbative approach are often visualized in a Maxwell-type diagram, as shown in Fig. III-2.

In the ensemble approach, the ground state (equilibrium) energy E is a function of N and a functional of the external potential $v(\mathbf{r})$, $E=E[N,v]$. The response of the system to changes of arguments is

$$dE = \left(\frac{\partial E[N,v]}{\partial N} \right)_v \mathrm{d}N + \int \left(\frac{\delta E[N,v]}{\delta v(\mathbf{r})} \right)_N \delta v(\mathbf{r}) \mathrm{d}\mathbf{r} \tag{3-14}$$

The first term defines the chemical potential

$$\mu[N,v] = \left(\frac{\partial E[N,v]}{\partial N} \right)_v , \tag{3-15}$$

while the second term, according to Eq.(2-49) generalized to the ensemble approach, gives

$$\left(\frac{\partial E[N,v]}{\partial v(\mathbf{r})} \right)_N = \rho(\mathbf{r};N,v). \tag{3-16}$$

Changes of μ are

$$d\mu[N,v] = \left(\frac{\partial \mu[N,v]}{\partial N} \right)_v dN + \int \left(\frac{\delta \mu[N,v]}{\delta v(\mathbf{r})} \right)_N \delta v(\mathbf{r}) \mathrm{d}\mathbf{r} . \tag{3-17}$$

The first term defines the global hardness[74]

$$\eta[N,v] = \left(\frac{\partial \mu[N,v]}{\partial N} \right)_v = \left(\frac{\partial^2 E[N,v]}{\partial^2 N} \right)_v = \frac{1}{S[N,v]}, \tag{3-18}$$

(here $S[N,v]$ is known as the global softness[75]). The global hardness can be viewed as a resistance of the system towards charge transfer. These global measures are used to express the Pearson's hard-soft and acid-base (HSAB) principle[76] that states hard acids preferably interact with hard bases, and soft acids with soft bases. The maximum hardness principle[77] seems to be a rule of nature. It states that molecules arrange themselves to be as hard as possible.

The second term defines the Fukui function

$$f(\mathbf{r};N,v) = \left(\frac{\delta \mu[N,v]}{\delta v(\mathbf{r})} \right)_N = \frac{\delta \partial E[N,v]}{\delta v(\mathbf{r}) \partial N} = \left(\frac{\partial \rho[\mathbf{r};N,v]}{\partial N} \right)_v . \tag{3-19}$$

Like the electron density, it is a local function. It expresses "the change of the electron density at each point **r** when the total number of electrons is changed". The local nature of the Fukui function provides information related to the reactivity at different sites (atoms, fragments) within a molecule. It can also be used as a factor governing the regioselectivity of chemical reactions.[78] This quantity can be viewed as a generalization of Fukui's frontier MO concept and plays a key role in linking Frontier MO Theory and the HSAB principle.

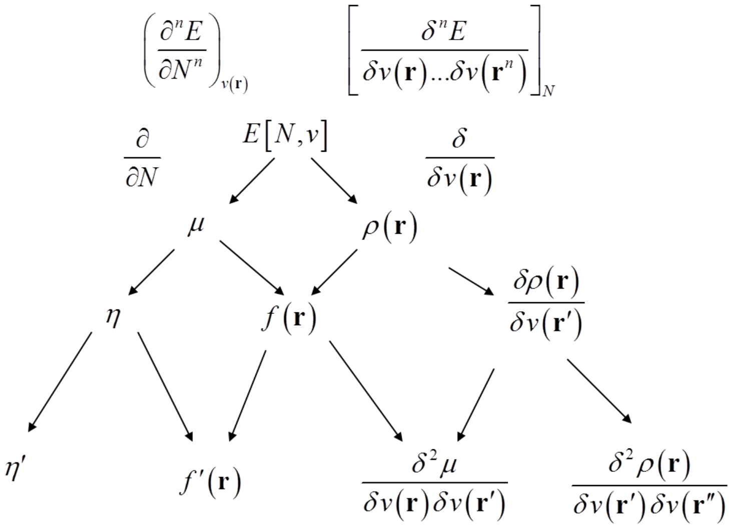
$\left(\frac{\partial^n E}{\partial N^n}\right)_{v(\mathbf{r})}$
$\left[\frac{\delta^n E}{\delta v(\mathbf{r})\ldots\delta v(\mathbf{r}^n)}\right]_N$
$\frac{\partial}{\partial N}$
$E[N,v]$
$\frac{\delta}{\delta v(\mathbf{r})}$
μ
$\rho(\mathbf{r})$
η
$f(\mathbf{r})$
$\frac{\delta\rho(\mathbf{r})}{\delta v(\mathbf{r}')}$
η'
$f'(\mathbf{r})$
$\frac{\delta^2\mu}{\delta v(\mathbf{r})\delta v(\mathbf{r}')}$
$\frac{\delta^2\rho(\mathbf{r})}{\delta v(\mathbf{r}')\delta v(\mathbf{r}'')}$

Fig. III-2 Maxwell scheme for conceptual DFT reactivity indices.

As follows from the ensemble description of the system, $E[N,v]$ at zero temperature limit is a piecewise-linear function of N. Therefore, $\mu[N,v]$ and $f(\mathbf{r};N,v)$ are discontinuous functions of N at integer points, constant between these points. Hence, at integer point N_0 one defines $f^+(\mathbf{r})$ as the derivative (Eq. (3-19)) taken from the side of $N > N_0$ and $f^-(\mathbf{r})$ – from the side of $N < N_0$. The functions $f^+(\mathbf{r})$and $f^-(\mathbf{r})$represent reactivity indices for nucleophilic and electrophilic attacks, respectively. Within Kohn–Sham theory, the $f^+(\mathbf{r})$ and $f^-(\mathbf{r})$ are associated with the LUMO frontier molecular orbital of the system and HOMO frontier molecular orbital of the system, respectively.[79] The $f^+(\mathbf{r})$ measures the reactivity towards an electron donor, while $f^-(\mathbf{r})$ measures the reactivity towards an electron acceptor.

In addition to these reactivity indices, other reactivity indices have been derived in the grand canonical ensemble as the derivatives with respect to the chemical potential and/or external potential, e.g. the chemical global and local softness.

Another quantity is the electrophilicity index, given in terms of μ and η as $\omega = \mu^2/2\eta$. It is a descriptor of reactivity that allows for a quantitative classification of the global electrophilic nature of a molecule within a relative scale, it is the power of a system to ‘soak up’ electrons.[80] It should be remarked that this quantity is, strictly speaking, not as an energy derivative. The intrinsic global electrophilicity has been associated to the lowering in energy predicted for any chemical species under interaction with a perfect electron donor (e.g.

an ideal zero chemical potential free electron environment). The size of the energy decrease is associated with the ability of the species to act as an electrophile.

Traditional Conceptual DFT indices typically model electron transfer and are thus not ideal when the transfer of electron spin is also important. Usually, changes in electron spin are coupled to electron transfer in radical reactions. However, in some cases, the extent of electron transfer is minimal with respect to changes in spin polarization. The need for a general framework allowing discussing chemical reactivity that includes both electron transfers, which is already incorporated in conventional Conceptual DFT, and spin polarization, which is not, motivated the development of spin-polarized versions of Conceptual DFT. The spin-polarized conceptual DFT formalism introduced by Galvan and co-workers[81,82] and a rigorous mathematical description given, for example, by Pérez, Chamorro and Ayers[83] insights useful chemical information. The introduction of new dual descriptors by E. Chamorro and coworkers[84] give further information about the molecular regions where the spin-polarization process linking states of different multiplicity will drive electron density and spin density changes.

III.C Alchemical derivatives

The definition of transmutation is extended beyond the historical definition of the transmutation as the conversion of base metals into gold or silver (the origin why these derivative are called alchemical). Here, transmutation means moving from one point to another in the conformational space at constant electron number ($dN=0$) and frozen geometry ($d\mathbf{R}^M=0$). The change in the total energy in the canonical ensemble, Eq.(3-8) can be rewritten as

$$\begin{aligned}
dW &= W\left[N,\mathbf{Z}^M+d\mathbf{Z}^M,\mathbf{R}^M\right]-W\left[N,\mathbf{Z}^M,\mathbf{R}^M\right] \\
&= \sum_A\left(\frac{\partial W}{\partial Z_A}\right)_{N,\mathbf{Z}_R^M,\mathbf{R}^M} dZ_A+\frac{1}{2}\sum_{AB}\left(\frac{\partial^2 W}{\partial Z_B\partial Z_A}\right)_{N,\mathbf{Z}_R^M,\mathbf{R}^M} dZ_A dZ_B+\ldots \\
&= \left\{\sum_A\left(\frac{\partial V_{\mathrm{nn}}}{\partial Z_A}\right)_{N,\mathbf{Z}_R^M,\mathbf{R}^M} dZ_A+\frac{1}{2}\sum_{AB}\left(\frac{\partial^2 V_{\mathrm{nn}}}{\partial Z_B\partial Z_A}\right)_{N,\mathbf{Z}_R^M,\mathbf{R}^M} dZ_A dZ_B\right\} \\
&\quad+\left[\begin{array}{l}\displaystyle\sum_A\left(\frac{\partial E}{\partial Z_A}\right)_{N,\mathbf{Z}_R^M,\mathbf{R}^M} dZ_A+\frac{1}{2}\sum_{AB}\left(\frac{\partial^2 E}{\partial Z_B\partial Z_A}\right)_{N,\mathbf{Z}_R^M,\mathbf{R}^M} dZ_A dZ_B \\ \displaystyle+\frac{1}{6}\sum_{ABC}\left(\frac{\partial^3 E}{\partial Z_C\partial Z_B\partial Z_A}\right)_{N,\mathbf{Z}_R^M,\mathbf{R}^M} dZ_A dZ_B dZ_C\end{array}\right],
\end{aligned} \tag{3-20}$$

where the subscript $\mathbf{Z}_R^M$ indicates that all other charges are kept fixed.

The terms in braces, $\{\ \}$, are the first- and the second-order terms resulting from V_{nn}:

$$\mu_A^{\mathrm{al,nuc}}\left[\mathbf{Z}^M,\mathbf{R}^M\right] \equiv \left(\frac{\partial V_{\mathrm{nn}}\left[\mathbf{Z}^M,\mathbf{R}^M\right]}{\partial Z_A}\right)_{\mathbf{Z}_R^M,\mathbf{R}^M} = \sum_{B\neq A}\frac{Z_B}{R_{AB}}, \tag{3-21}$$

$$\eta_{AB}^{\mathrm{al,nuc}}\left[\mathbf{Z}^M,\mathbf{R}^M\right] \equiv \left(\frac{\partial^2 V_{\mathrm{nn}}\left[\mathbf{Z}^M,\mathbf{R}^M\right]}{\partial Z_B \partial Z_A}\right)_{\mathbf{Z}_R^M,\mathbf{R}^M} = \left(1-\delta_{AB}\right)\frac{1}{R_{AB}} = \eta_{BA}^{\mathrm{al,nuc}}, \tag{3-22}$$

where $R_{AB} = \left|\mathbf{R}_A - \mathbf{R}_B\right|$. The higher-order derivatives are trivially equal to zero. The terms in brackets, $[\ ]$ are derivatives of the electronic energy, which depend on $\mathbf{Z}^M$ via the external potential, $v\left(\mathbf{r},\mathbf{Z}^M,\mathbf{R}^M\right)$. The electronic contribution to the electrostatic potential at site A (the nuclear-electron attraction energy per unit nuclear charge,[85] also involved in the diamagnetic shielding of an atom in a molecule[86]) and the effective 'screened' potential (hardness) defined by the linear response kernel are

$$\begin{aligned}\mu_A^{\mathrm{al,el}}\left[N,\mathbf{Z}^M,\mathbf{R}^M\right] &\equiv \left(\frac{\partial E}{\partial Z_A}\right)_{N,\mathbf{Z}_R^M,\mathbf{R}^M} = \int\left(\frac{\delta E[N,v]}{\delta v(\mathbf{r})}\right)_N \left(\frac{\partial v(\mathbf{r})}{\partial Z_A}\right)_{\mathbf{Z}_R^M,\mathbf{R}^M} d\mathbf{r} \\ &= -\int \frac{\rho(\mathbf{r})}{\left|\mathbf{r}-\mathbf{R}_A\right|} d\mathbf{r},\end{aligned} \tag{3-23}$$

$$\begin{aligned}\eta_{AB}^{\mathrm{al,el}}\left[N,\mathbf{Z}^M,\mathbf{R}^M\right] &\equiv \left(\frac{\partial^2 E}{\partial Z_B \partial Z_A}\right)_{N,\mathbf{Z}_R^M,\mathbf{R}^M} \\ &= \iint\left(\frac{\delta E[N,v]}{\delta v(\mathbf{r})\delta v(\mathbf{r}')}\right)_N \left(\frac{\partial v(\mathbf{r})}{\partial Z_A}\right)_{\mathbf{Z}_R^M,\mathbf{R}^M} \left(\frac{\partial v(\mathbf{r}')}{\partial Z_B}\right)_{\mathbf{Z}_R^M,\mathbf{R}^M} d\mathbf{r}d\mathbf{r}' \\ &= \iint \frac{\chi(\mathbf{r},\mathbf{r}')}{\left|\mathbf{r}-\mathbf{R}_A\right|\left|\mathbf{r}-\mathbf{R}_B\right|} d\mathbf{r}d\mathbf{r}' = \eta_{BA}^{\mathrm{al,el}}.\end{aligned} \tag{3-24}$$

Here, the relation $\left(\delta E[N,v]/\delta v(\mathbf{r})\right)_N = \rho(\mathbf{r};N,v)$ is used and the linear-response function $\chi(\mathbf{r},\mathbf{r}')$ has been introduced

$$\chi(\mathbf{r},\mathbf{r}') \equiv \left(\frac{\delta^2 E[N,v]}{\delta v(\mathbf{r})\delta v(\mathbf{r}')}\right)_N = \left(\frac{\delta\rho(\mathbf{r})}{\delta v(\mathbf{r}')}\right)_N. \tag{3-25}$$

Because the linear response function is a symmetric kernel, the effective potential, Eq.(3-24), is symmetric in AB. Using the linear dependence of $v(\mathbf{r})$ on the nuclear charge:

$$\left(\frac{\partial v\left(\mathbf{r},\mathbf{Z}^{M},\mathbf{R}^{M}\right)}{\partial Z_{A}}\right)_{\mathbf{Z}_{R}^{M},\mathbf{R}^{M}}=-\frac{1}{\left|\mathbf{r}-\mathbf{R}_{A}\right|}, \tag{3-26}$$

the higher-order derivatives of the electronic energy can be written as

$$\left(\frac{\partial^{k}E}{\partial Z_{k}...\partial Z_{1}}\right)_{N,\mathbf{Z}_{R}^{M},\mathbf{R}^{M}}=\left(-1\right)^{k}\int...\int\frac{\chi\left(\mathbf{r}_{1},..,\mathbf{r}_{k}\right)}{\left|\mathbf{r}-\mathbf{R}_{1}\right|...\left|\mathbf{r}-\mathbf{R}_{k}\right|}d\mathbf{r}_{1}...d\mathbf{r}_{k},\quad Z_{1},...,Z_{k}\subset\mathbf{Z}^{M}, \tag{3-27}$$

where $\chi\left(\mathbf{r}_{1},..,\mathbf{r}_{k}\right)\equiv\left(\delta^{k-1}\rho\left(\mathbf{r}_{1}\right)/\delta v\left(\mathbf{r}_{2}\right)...\delta v\left(\mathbf{r}_{k}\right)\right)_{N}$ is the functional derivative of the density with respect to the perturbation potential, evaluated at the ground-state density value for a constant particle number.

Finally, the alchemical derivatives in their global version are (the subscripts 'al' means alchemical)

- the alchemical potential of nucleus at position $\mathbf{R}_{A}$ [35,52]

$$\mu_{A}^{\text{al}}\left[N,\mathbf{Z}^{M},\mathbf{R}^{M}\right]\equiv\left(\frac{\partial W}{\partial Z_{A}}\right)_{N,\mathbf{Z}_{R}^{M},\mathbf{R}^{M}}=\mu_{A}^{\text{al,el}}+\mu_{A}^{\text{al,nuc}}, \tag{3-28}$$

which measures the transmutational tendency for the atom at position $\mathbf{R}_{A}$ in the molecule. If there is a vacancy at this position, it is related to the proton affinity.

- the alchemical hardness[35]

$$\eta_{AB}^{\text{al}}\left[N,\mathbf{Z}^{M},\mathbf{R}^{M}\right]\equiv\left(\frac{\partial^{2}W}{\partial Z_{A}\partial Z_{B}}\right)_{N,\mathbf{Z}_{R}^{M},\mathbf{R}^{M}}=\eta_{AB}^{\text{al,el}}+\eta_{AB}^{\text{al,nuc}}=\eta_{BA}^{\text{al}}, \tag{3-29}$$

which measures the transmutational resistivity for the atoms at the positions $\mathbf{R}_{A}$ and $\mathbf{R}_{B}$ in the molecule. The alchemical hardness η_{AB}^{al} is simply the effective 'screened' potential.

- the alchemical stiffness:

$$\begin{aligned}\sigma_{ABC}^{\text{al}}\left[N,\mathbf{Z}^{M},\mathbf{R}^{M}\right]\equiv\left(\frac{\partial^{3}W}{\partial Z_{A}\partial Z_{B}\partial Z_{C}}\right)_{N,\mathbf{Z}^{M},\mathbf{R}^{M}}&=\left(\frac{\partial^{3}E}{\partial Z_{A}\partial Z_{B}\partial Z_{C}}\right)_{N,\mathbf{Z}^{M},\mathbf{R}^{M}}\\&=-\iiint\frac{\chi\left(\mathbf{r}_{1},\mathbf{r}_{2},\mathbf{r}_{3}\right)}{\left|\mathbf{r}_{1}-\mathbf{R}_{A}\right|\left|\mathbf{r}_{2}-\mathbf{R}_{B}\right|\left|\mathbf{r}_{3}-\mathbf{R}_{C}\right|}d\mathbf{r}_{1}d\mathbf{r}_{2}d\mathbf{r}_{3}=\sigma_{ABC}^{\text{al,el}}\end{aligned} \tag{3-30}$$

and so on for the higher-order derivatives, $\left(\partial^{k}W/\partial Z_{k}...\partial Z_{1}\right)_{N,\mathbf{Z}_{R},\mathbf{R}}$. These higher-order terms are usually assumed to be qualitatively unimportant; although there is a growing realization that the "chemical perturbations" are sometimes too large for this to be true.[55,57]

III.D The alchemical transmutation energy

To illustrate the usefulness of the alchemical derivatives in exploring chemical space, the deprotonation of many molecules, the transmutation of the nitrogen molecule, and the substitution of isoelectronic B-N units for C-C units and N units for C-H units were chosen as examples of increasing complexity in the number of sites which are transmuted (will be discussed in next chapter). To preserve the predictive power of the Taylor expansion and to preserve a realistic perspective from a chemical point of view, only molecules at their equilibrium geometries are used as the basis for an expansion. Moreover, changes in the atomic charges are limited, $dZ_i \in \{-1,0,1\}$. The $\left(\mathbf{R}^M,\mathbf{Z}^M + d\mathbf{Z}^M\right)$ point in the conformational space will be named a transmutant of the molecule described by the $\left(\mathbf{R}^M,\mathbf{Z}^M\right)$ point in this space (see Fig. III-3). The *alchemical transmutation energy* connected with the displacement from $\left(\mathbf{R}^M,\mathbf{Z}^M\right)$ to $\left(\mathbf{R}^M,\mathbf{Z}^M + d\mathbf{Z}^M\right)$ is defined as

$$\begin{aligned} D^{\text{al}}\left[\mathbf{R}^M,\mathbf{Z}^M,d\mathbf{Z}^M\right] &\equiv \sum_i \mu_i^{\text{al}}\left[\mathbf{R}^M,\mathbf{Z}^M\right]dZ_i^M + \frac{1}{2}\sum_{ij}\eta_{ij}^{\text{al}}\left[\mathbf{R}^M,\mathbf{Z}^M\right]dZ_i^M dZ_j^M \\ &= D^{\text{al},\mu}\left[\mathbf{R}^M,\mathbf{Z}^M,d\mathbf{Z}^M\right] + D^{\text{al},\eta}\left[\mathbf{R}^M,\mathbf{Z}^M,d\mathbf{Z}^M\right], \end{aligned} \qquad (3\text{-}31)$$

where $D^{\text{al},\mu}$ and $D^{\text{al},\eta}$ are the contributions from the first and second derivatives, respectively.

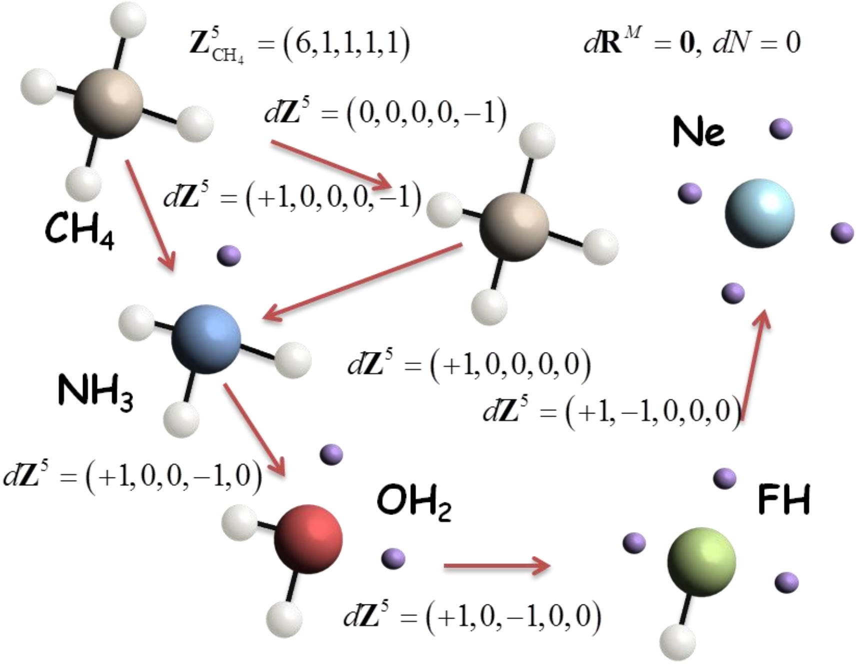
$\mathbf{Z}^5_{\text{CH}_4} = (6,1,1,1,1)$
$d\mathbf{R}^M = \mathbf{0},\ dN = 0$
$d\mathbf{Z}^5 = (0,0,0,0,-1)$
Ne
$d\mathbf{Z}^5 = (+1,0,0,0,-1)$
CH₄
NH₃
$d\mathbf{Z}^5 = (+1,0,0,0,0)$
$d\mathbf{Z}^5 = (+1,-1,0,0,0)$
$d\mathbf{Z}^5 = (+1,0,0,-1,0)$
OH₂
FH
$d\mathbf{Z}^5 = (+1,0,-1,0,0)$

Fig. III-3. Transmutation "tour" from methyl to neon atom.

The energy can be decomposed also into the electronic and the nuclear parts:

$$\begin{aligned} D^{\mathrm{al}}\left[\mathbf{R}^M,\mathbf{Z}^M,d\mathbf{Z}^M\right] &= D^{\mathrm{al,el}}\left[\mathbf{R}^M,\mathbf{Z}^M,d\mathbf{Z}^M\right]+D^{\mathrm{al,nuc}}\left[\mathbf{R}^M,\mathbf{Z}^M,d\mathbf{Z}^M\right] \\ &=\left(\sum_i \mu_i^{\mathrm{al,el}}\left[\mathbf{R}^M,\mathbf{Z}^M\right]dZ_i^M+\frac{1}{2}\sum_{ij}\eta_{ij}^{\mathrm{al,el}}\left[\mathbf{R}^M,\mathbf{Z}^M\right]dZ_i^M dZ_j^M\right) \\ &+\left(\sum_i \mu_i^{\mathrm{al,nuc}}\left[\mathbf{R}^M,\mathbf{Z}^M\right]dZ_i^M+\frac{1}{2}\sum_{ij}\eta_{ij}^{\mathrm{al,nuc}}\left[\mathbf{R}^M,\mathbf{Z}^M\right]dZ_i^M dZ_j^M\right) \\ &=\left(D^{\mathrm{al},\mu,\mathrm{el}}+D^{\mathrm{al},\mu,\mathrm{nuc}}\right)+\left(D^{\mathrm{al},\eta,\mathrm{el}}+D^{\mathrm{al},\eta,\mathrm{nuc}}\right) \end{aligned} \qquad (3\text{-}32)$$

where the $D^{\mathrm{al},\mu}$ and $D^{\mathrm{al},\eta}$ are the contribution from the first- and second-order derivatives, respectively. The component resulting from the wavefunction modification, $D^{\mathrm{al},\eta,\mathrm{el}}$ is calculated using the coupled perturbed self consistent field theory (see Ref.[87]) for computing details and basis set and weights dependencies on the nuclear charge).

To have a clear understanding of alchemical transmutation energy, let us consider a diatomic molecule AB in the reaction type

$$AB \rightleftarrows CD\,. \qquad (3\text{-}33)$$

The vertical (transmutation) energy is (@ means at equilibrium geometry of)

$$D_{AB\to CD}^{\mathrm{ver},@AB}=W_{CD}^{@AB}-W_{AB}^{@AB}=\left(E_{CD}^{@AB}-E_{AB}^{@AB}\right)+\left(V_{CD}^{@AB}-V_{AB}^{@AB}\right)=-D_{CD\to AB}^{\mathrm{ver},@AB}\,, \qquad (3\text{-}34)$$

where the W, E and V are the energies defined in Eq. (3-6). Note that calculation of this vertical energy needs two runs of Kohn-Sham program: for CD and AB molecule. This vertical energy can be approximated by the alchemical derivatives (limited to second-order Taylor expansion) as (see Eq.(3-32))

$$D_{AB\to CD}^{\mathrm{ver}@AB}\approx D_{AB\to CD}^{\mathrm{al}@AB}=\left(\begin{aligned}&\left[\begin{aligned}&\left(\mu_A^{\mathrm{al,el}}dZ_A+\mu_B^{\mathrm{al,el}}dZ_B\right)\\&+\frac{1}{2}\left(\eta_{AA}^{\mathrm{al,el}}dZ_A^2+2\eta_{AB}^{\mathrm{al,el}}dZ_A dZ_B+\eta_{BB}^{\mathrm{al,el}}dZ_B^2\right)\end{aligned}\right]\\&\left\{\begin{aligned}&+\mu_A^{\mathrm{al,nuc}}dZ_A+\mu_B^{\mathrm{al,nuc}}dZ_B\\&+\frac{1}{2}\left(\eta_{AA}^{\mathrm{al,nuc}}dZ_A^2+2\eta_{AB}^{\mathrm{al,nuc}}dZ_A dZ_B+\eta_{BB}^{\mathrm{al,nuc}}dZ_B^2\right)\end{aligned}\right\}\end{aligned}\right) \qquad (3\text{-}35)$$

Hence, the nuclear contribution to the transmutation energy will be

$$\begin{aligned} D_{AB\to CD}^{\mathrm{al,nuc}}&=\frac{1}{R_{AB}}\left\{\left(Z_B dZ_A+Z_A dZ_B\right)+\left(1-\delta_{AB}\right)dZ_A dZ_B\right\}\\ &=\frac{1}{R_{AB}}\left\{\left(Z_B dZ_A+Z_A dZ_B\right)+dZ_A dZ_B\right\}\end{aligned} \qquad (3\text{-}36)$$

where the last equality is for $A\neq B$.

If proton (charge) number N_p is kept constant during transmutation $\left(dZ_A = -dZ_B\right)$, Eq.(3-36) becomes

$$D_{AB\to CD}^{\text{al,nuc}} = \frac{1}{R_{AB}}\left\{\left(Z_B - Z_A\right)dZ_A - dZ_A^2\right\} \tag{3-37}$$

while the electronic part is

$$D_{AB\to CD}^{\text{al,el}} = \left(\mu_A^{al,\text{el}} - \mu_B^{al,\text{el}}\right)dZ_A + \frac{1}{2}\left(\eta_{AA}^{al,\text{el}} - 2\eta_{AB}^{al,\text{el}} + \eta_{BB}^{al,\text{el}}\right)dZ_A^2. \tag{3-38}$$

On the other hand, if we consider a transmutation type $AB \rightleftarrows BA$ it leads to

$$D_{AB\to BA}^{\text{al,nuc}} = 0, \qquad D_{AB\to BA}^{\text{al,el}} = 0 \tag{3-39}$$

For a homonuclear diatomic molecule of transmutation type $A_1A_2 \to CD$ with $dZ_{A_2} = -dZ_{A_1}$, we have

$$D_{A_1A_2\to CD}^{\text{al,nuc}} = -\frac{dZ_{A_1}^2}{R_{A_1A_2}}. \tag{3-40}$$

The electronic contribution of such type transmutation will be

$$D_{A_1A_2\to CD}^{\text{al,el}} = \left(\mu_{A_1}^{\text{al,el}} - \mu_{A_2}^{\text{al,el}}\right)dZ_{A_1} + \frac{1}{2}\left(\eta_{A_1A_1}^{\text{al,el}} - 2\eta_{A_1A_2}^{\text{al,el}} + \eta_{A_2A_2}^{\text{al,el}}\right)dZ_{A_1}^2 = \left(\eta_{A_1A_1}^{\text{al,el}} - \eta_{A_1A_2}^{\text{al,el}}\right)dZ_{A_1}^2, \tag{3-41}$$

with $\mu_{A_1}^{\text{al,el}} = \mu_{A_2}^{\text{al,el}}$ and $\eta_{A_1A_1}^{\text{al,el}} = \eta_{A_2A_2}^{\text{al,el}}$. For reverse transmutation $CD \to A_1A_2$ with $dZ_C = -dZ_D$, we find

$$D_{CD\to A_1A_2}^{\text{al,nuc}} = \frac{1}{R_{CD}}\left\{\left(Z_D - Z_C\right)dZ_C - dZ_C^2\right\}, \tag{3-42}$$

$$D_{CD\to A_1A_2}^{\text{al,el}} = \left(\mu_C^{\text{al,el}} - \mu_D^{\text{al,el}}\right)dZ_C + \frac{1}{2}\left(\eta_{CC}^{\text{al,el}} - 2\eta_{CD}^{\text{al,el}} + \eta_{DD}^{\text{al,el}}\right)dZ_C^2 \tag{3-43}$$

In general, according to Eq. (3-34)we have

$$D_{AB\to CD}^{\text{ver,@}AB} + D_{CD\to AB}^{\text{ver,@}AB} = 0. \tag{3-44}$$

The errors between the vertical and alchemical transmutation energies for the forward and reverse reactions are

$$\Delta D_{AB\to CD}^{\text{al,@}AB} = D_{AB\to CD}^{\text{ver,@}AB} - D_{AB\to CD}^{\text{al,@}AB} \quad \text{and} \quad \Delta D_{CD\to AB}^{\text{al,@}AB} = D_{CD\to AB}^{\text{ver,@}AB} - D_{CD\to AB}^{\text{al,@}AB}. \tag{3-45}$$

The difference between the vertical and alchemical transmutation energies includes also an error connected with the omission of higher-order terms in the Taylor expansion.

Chapter IV. Examples and Applications of Alchemical Derivatives

IV.A Deprotonation energy.

In analogy to the vertical ionization energy, the energy change corresponding to an ionization reaction leading to formation of the ion in a configuration which is the same as the equilibrium geometry of the neutral molecule at the ground state, the *vertical deprotonation energy* is defined as the energy change associated with the deprotonation reaction

$$\mathrm{AH} \rightarrow \mathrm{A}^{-}+\mathrm{H}^{+} \qquad D_{\mathrm{AH}}^{\mathrm{ver},@AH} = W_{\mathrm{A}^{-}}^{@AH} - W_{\mathrm{AH}}^{@AH}, \tag{4-1}$$

where $W_{\mathrm{AH}}^{@\mathrm{AH}}$, $W_{\mathrm{A}^{-}}^{@\mathrm{AH}}$ are the total energies of the acid and its conjugate base at the equilibrium geometry of the acid, respectively. D^{ver} is a quantity defined at $0\mathrm{K}$, and is therefore not strictly related to the experimental deprotonation energy, which is the enthalpy change of this reaction. The experimental deprotonation energy can be approximated by *ab-initio* calculations at room temperature as follows

$$D_{\mathrm{AH}}^{\exp} \approx D_{\mathrm{AH}}^{0} \equiv D_{\mathrm{AH}}^{\mathrm{ver},@\mathrm{AH}} + \Delta W_{\mathrm{A}^{-}}^{\mathrm{rel}} + \Delta H_{\mathrm{AH}}^{\mathrm{A}^{-}} + H_{\mathrm{corr}}^{\mathrm{H}^{+}}, \tag{4-2}$$

where $\Delta W_{\mathrm{A}^{-}}^{\mathrm{rel}} = \left(W_{\mathrm{A}^{-}}^{@\mathrm{A}^{-}} - W_{\mathrm{A}^{-}}^{@\mathrm{AH}}\right)$ is the geometry relaxation correction, $\Delta H_{\mathrm{AH}}^{\mathrm{A}^{-}} = \left(H_{\mathrm{corr}}^{\mathrm{A}^{-}} - H_{\mathrm{corr}}^{\mathrm{AH}}\right)$ is the enthalpy correction and $H_{\mathrm{corr}}^{\mathrm{H}^{+}}$ is the calculated gas-phase enthalpy of the proton at room temperature.

The vertical deprotonation energy can be approximated using the Taylor expansion. For the M-atomic system with m hydrogens ($1 \le m \le M$) and the nuclear charge vector, $\mathbf{Z}_{\mathrm{AH}_m}^{M} = \left(1,...,1,Z_{m+1},...,Z_M\right)$, the annihilation of the proton at position $\mathbf{R}_i$ can be described by the transmutation vector

$$d\mathbf{Z}_{\mathrm{AH}_m,i}^{M} = \left(0,...,dZ_i = -1,0,...\right), \quad 1 \le i \le m, \tag{4-3}$$

The *alchemical deprotonation energy* is defined through the second order Taylor expansion

$$\mathrm{D}_{\mathrm{AH}_m,i}^{\mathrm{al}} = -\mu_i^{\mathrm{al}} + \frac{1}{2}\eta_{ii}^{\mathrm{al}} = \left(-\mu_i^{\mathrm{al,el}} + \frac{1}{2}\eta_{ii}^{\mathrm{al,el}}\right) - \mu_i^{\mathrm{al,nuc}}, \tag{4-4}$$

and can be used as the approximation to the vertical deprotonation energy. For a system with more than one hydrogen, a set of conjugate bases can be created. The conjugate base with the lowest $\mathrm{D}_{\mathrm{AH}_m,i}^{\mathrm{al}}$ is assumed to be the most stable one. The nuclear part of the alchemical potential always takes a positive value for the molecule, Eq.(3-21), and the hydrogen site with the highest value of $\mu_i^{\text{al,nuc}}$ can be put forward as the most probable deprotonation site. A simple example is the methylammonium cation $\text{CH}_3\text{-}\overset{\oplus}{\text{N}}\text{H}_3$, at the ethane geometry. A calculation shows that $\mu_{\text{H(-N)}}^{\text{al,nuc}} - \mu_{\text{H(-C)}}^{\text{al,nuc}} = 1/R_{\text{C-H}} > 0$ so that the deprotonation of the hydrogen attached to nitrogen atom produces the most stable anion as expected intuitively: $\text{CH}_3\text{-NH}_2$ is expected to be more stable than the dipolar $\overset{\ominus}{\text{C}}\text{H}_2\text{-}\overset{\oplus}{\text{N}}\text{H}_3$ structure. Another example, the *n*-propane deprotonation which yields two carbanions, shows that the electronic contribution is also important. The $\mu_i^{\text{al,nuc}}$ for the hydrogens connected to the central carbon is higher than for the terminal hydrogens. But, the most stable anion is predicted to be the 1-propyl anion rather than the isopropyl anion in agreement with the experimental results[88] and common knowledge.

In general, the one-site deprotonation, Eq.(4-1), can be extended to a *n*-tuple deprotonation reaction:

$$\text{AH}_m \rightarrow \left(\text{AH}_{m-n}\right)^{n-} + n\text{H}^+ . \tag{4-5}$$

This hypothetical gas-phase reaction does not precisely describe at which positions the protons were annihilated, but all conjugate bases of the AH_m acid can be represented by a set of transmutation vectors

$$\mathcal{D}_{\text{AH}_m}^n \equiv \left\{ d\mathbf{Z}_{\text{AH}_m}^M = \left(dZ_1,\cdots,dZ_m,0,\cdots\right) \middle| dZ_i \in \left\{-1,0\right\}, 1 \le i \le m, \sum_{i=1}^m dZ_m = -n \right\} . \tag{4-6}$$

For each of these vectors, the alchemical transmutation energy is

$$\text{D}_{\text{AH}_m}^{\text{al}}\left[d\mathbf{Z}_{\text{AH}_m}^M\right] \equiv \sum_{i=1}^n \mu_i^{\text{al}} dZ_i + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \eta_{ij}^{\text{al}} dZ_i dZ_j . \tag{4-7}$$

The minimizer of the alchemical transmutation energy can be viewed as the transmutation vector which leads to the more stable conjugate base, $\left(\text{AH}_{m-n}\right)^{n-}$.

The usefulness of the alchemical derivatives as chemical reactivity indices in the deprotonation energy was tested on the set of 20 molecules with one and a few hydrogen atoms (see Scheme IV-1). Two exchange-correlation functionals: B3LYP[89,90] and PBE[91] were used in a combination with the correlation consistent double- and triple-ζ basis sets proposed by Dunning and coworkers(along with their augmented counterparts).[11,92-94] The correlation coefficients and the mean absolute errors (MAEs) are shown in Table IV-1 (detailed data for these correlations are available from *Supporting Information Disc* (SID)).

$$X-H \rightarrow X^{-} + \mathrm{H}^{+}$$
$$X \in \left\{ \begin{array}{l} \mathrm{H, CH_3, CH_2CH_3, CHCH_2, CHO, COOCH_3, CN, NH_{2,}} \\ \mathrm{OCH_{3,}\ OH, F, CH_2F, CHF_2, CF_3, SiH_3, PH_2, SH, Cl, CH_2Cl, CCl_3} \end{array} \right\}$$

Scheme IV-1. Deprotonation of X-H molecule.

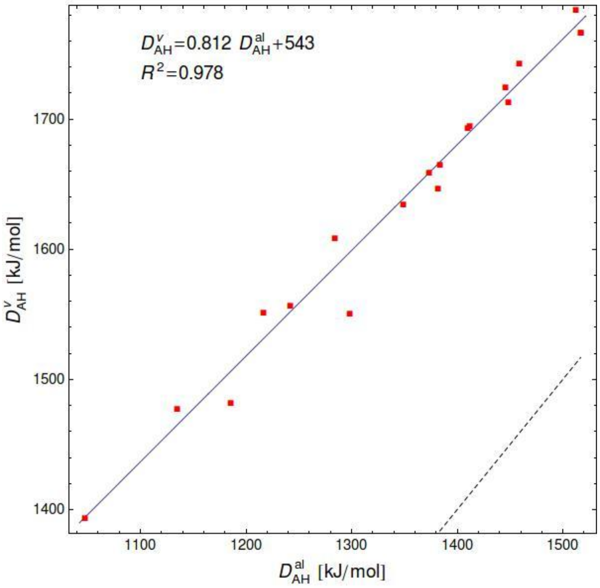

Fig. IV-1. Correlation of the alchemical deprotonation energy, Eq.(4-4), with the vertical deprotonation energy, Eq.(4-1), at B3LYP/aug-cc-pVDZ level. All values in KJ/mol, the dashed line is an exact result line

We conclude that functional dependencies are insignificant since the results for both functionals used in this work are very similar (the B3LYP results are slightly better than the PBE results if we consider the correlation coefficients and conversely if the MAEs are considered). Increasing the ζ splitting slightly decreases the correlation coefficient (and the increasing the MAE) for the $\mathrm{D}^{\mathrm{ver}}_{\mathrm{XH}}$ vs. $D^{\mathrm{al}}_{\mathrm{XH}}$ correlation. This can be explained by the presence of the artificial binding of the electron by the basis set. The same effect of the basis set is observed in the case of the electron affinity calculation.[95] However, contrary to the latter, the addition of diffuse functions improves the correlations with decreasing MAE (see Fig. IV-1). Based on the correlations between $\mathrm{D}^{\mathrm{exp}}_{\mathrm{XH}}$ and $\mathrm{D}^{0}_{\mathrm{XH}}$ (Table IV-1) we can explain that this is connected with the fact that the augmented basis sets most closely reproduce the experimental geometry. A similar effect was observed in the case of silicaceous materials,[96] for example. It is pertinent to comment on the possible influence of the fact that the basis set centered on the proton is still included in the conjugated base basis set after the annihilation of the proton. The calculations of the vertical deprotonation energy with the ghost atom at the deprotonation site barely change the results compared to the no-ghost atom calculations (the average difference is less than 10 kJ/mol). The most important conclusion is the very good correlation between $\mathrm{D}^{\mathrm{ver}}_{\mathrm{XH}}$ and $D^{\mathrm{al}}_{\mathrm{XH}}$ (e.g. B3LYP/aug-cc-pVDZ, R^2=0.978) indicating that with a simple alchemical type calculation (requiring only one single point calculation) trends in deprotonation energies can be retrieved. Note that when comparing D_{XH}^{exp} and D_{XH}^{al} contributions from geometry relaxation and enthalpy correction are to be included to fully reproduce the experimental data. Nevertheless, the alchemical derivative value is a trustworthy indicator of the experimental deprotonation energies in comparative studies as usually performed when exploring CS.

Table IV-1. Correlation coefficients and mean absolute error (in parentheses, in kJ/mol) for deprotonation energy for 20 X-H molecules. Comparison between experimental, vertical and alchemical values is made. Experimental data from http://webbook.nist.gov.

Functional	Basis	D_{XH}^{exp} vs.			D_{XH}^{ver} vs. D_{XH}^{al}
		D_{XH}^{o}	D_{XH}^{ver}	D_{XH}^{al}	
B3LYP	cc-pVDZ	0.859 (77)	0.851 (138)	0.804 (181)	0.973 (319)
	cc-pVTZ	0.941 (41)	0.893 (94)	0.849 (260)	0.969 (354)
	aug -cc-pVDZ	0.964 (19)	0898 (40)	0.895 (250)	0.978 (290)
	aug -cc-pVTZ	0.961 (13)	0.889 (44)	0.878 (279)	0.954 (324)
PBE	cc-pVDZ	0.825 (74)	0.829 (134)	0.783 (181)	0.972 (315)
	cc-pVTZ	0.919 (37)	0.884 (88)	0.832 (249)	0.965 (352)
	aug -cc-pVDZ*	0.954 (24)	0.901 (33)	0.891 (253)	0.975 (284)
	aug -cc-pVTZ	0.950 (20)	0.891 (35)	0.869 (284)	0.946 (319)

* H_2 is not included in correlation due to numerical instability

IV.B Nitrogen molecule case.

As the next example, we consider the transmutation of the nitrogen molecule. Five chemically relevant transmutations of the nitrogen molecule are considered in this work (see Scheme IV-2). The nitrogen molecule and its “transmutants” are examples of the isosters[97] and isoelectronic molecules. All these transmutants can be expected to exist in the atmosphere of the terrestrial planets which have a dense atmosphere of N_2.[98] C_2^{2-} is one of the few doubly negative ions to have bound valence states,[99] and has been detected in comets.[100] O_2^{2+} is kinetically stable and it is considered as the molecule with the shortest bond between any two heavy atoms.[101] The nitrosonium cation is the key species in the process of nitrosation, an important process in the cell biochemistry.[102] The cyanide ion and carbon monoxide are well known in organic and inorganic chemistry, both as reagent and ligand.

$$\begin{tikzcd}
 & CO & NO^{+} & O_2^{2+} \\
dZ_2 & CN^{-} & N_2 \arrow[ul, leftrightarrow] \arrow[u, leftrightarrow] \arrow[ur, leftrightarrow] \arrow[l, leftrightarrow] \arrow[r, leftrightarrow] \arrow[dl, leftrightarrow] \arrow[d, leftrightarrow] \arrow[dr, leftrightarrow] & ON^{+} \\
 & C_2^{2-} & NC^{-} & OC \\
 & & dZ_1 &
\end{tikzcd}$$

Scheme IV-2 Transmutation of the nitrogen molecule with possible changes of charge at the nuclear sites $dZ_1, dZ_2 \in \{-1,0,1\}$.

The set of the transmutation vectors for the nitrogen transmutation is simply

$$\mathcal{D}_{N_2} \equiv \left\{ d\mathbf{Z}_{N_2}^{M} = (dZ_1, dZ_2) \middle| dZ_i \in \{-1,0,1\} \right\} \tag{4-8}$$

and the alchemical transmutation energies are (the fact that $\mu_1^{\mathrm{al},N_2} = \mu_2^{\mathrm{al},N_2}$, $\eta_{1,2}^{\mathrm{al},N_2} = \eta_{2,1}^{\mathrm{al},N_2}$ and $\eta_{1,1}^{\mathrm{al},N_2} = \eta_{2,2}^{\mathrm{al},N_2}$ is used)

$$\begin{aligned} D_{N_2 \to CO}^{\mathrm{al},N_2} = {} & \left(\mu_1^{\mathrm{al},N_2} + \frac{1}{2} \eta_{1,1}^{\mathrm{al},N_2} (dZ_1 + dZ_2) \right) (dZ_1 + dZ_2) \\ & + \left(\eta_{1,2}^{\mathrm{al},N_2} - \eta_{1,1}^{\mathrm{al},N_2} \right) dZ_1 dZ_2 \end{aligned} \tag*{.(4-9)}$$

In the case of the $N_2 \to CO$ transmutation, the above equation becomes simply

$$D_{N_2 \to CO}^{\mathrm{al},N_2} = \eta_{1,1}^{\mathrm{al},N_2} - \eta_{1,2}^{\mathrm{al},N_2}, \tag{4-10}$$

due to $dZ_1 + dZ_2 = 0$, $dZ_2 = 1$. For the reverse transmutation, $CO \to N_2$ at the nitrogen molecule geometry, the alchemical transmutation energy is

$$D_{CO \to N_2}^{\mathrm{al},N_2} = \left(\mu_O^{\mathrm{al},N_2} - \mu_C^{\mathrm{al},N_2} \right) + \frac{1}{2} \left(\eta_{O,O}^{\mathrm{al},N_2} + \eta_{C,C}^{\mathrm{al},N_2} - 2\eta_{C,O}^{\mathrm{al},N_2} \right). \tag{4-11}$$

The difference between the vertical and alchemical transmutation energies, $\Delta D_{N_2 \to CO}^{\mathrm{al},N_2} = D_{N_2 \to CO}^{\mathrm{ver},N_2} - D_{N_2 \to CO}^{\mathrm{al},N_2}$ and $\Delta D_{CO \to N_2}^{\mathrm{al},N_2} = -D_{N_2 \to CO}^{\mathrm{ver},N_2} - D_{CO \to N_2}^{\mathrm{al},N_2}$ includes an error connected with the omission of higher-order terms in the Taylor expansion. In case of the $N_2 \rightarrow CO$ transmutation, only the even order terms survive due to the cancellation of the odd terms.

Table IV-2. The mean absolute error between the vertical and the alchemical transmutation energy for all 5 nitrogen molecule transmutations (Scheme IV-2 Transmutation of the nitrogen molecule with possible changes of charge at the nuclear sites $dZ_1, dZ_2 \in \{-1,0,1\}$..). The MAE for $D^{\mathrm{ver},N_2}_{N_2\rightarrow AB}$ vs. $D^{\mathrm{al},N_2}_{N_2\rightarrow AB}$ is tabulated, and the values for $D^{\mathrm{ver},N_2}_{AB\rightarrow N_2}$ vs. $D^{\mathrm{al},N_2}_{AB\rightarrow N_2}$ are given in parentheses. All data in a.u.

	B3LYP		PBE	
	n		n	
	D	T	D	T
cc-pVnZ	1.148 (1.125)	0.275 (0.266)	1.142 (1.119)	0.279 (0.266)
aug-cc-pVnZ	0.755 (0.715)	0.158 (0.122)	0.751 (0.711)	0.162 (0.121)
cc-pCVnZ	0.115 (0.121)	**0.034** (**0.025**)	0.113 (0.119)	**0.034** (**0.023**)
aug-cc-pCVnZ	0.096 (0.079)	0.041 (0.047)	0.095 (0.078)	0.042 (0.045)

For testing the usefulness of the alchemical derivatives in the case of such transmutations, the same pair of the exchange-correlation functionals as in Sec.IV.A was used. The basis set was expanded by inclusion of the additional tight functions (large exponents) in order to recover the core-core and the core-valence correlation[103] which is expected to be significant in such transmutations. In Table IV-2, the mean absolute errors for the $D^{ver,\mathrm{N}_2}_{\mathrm{N}_2\rightarrow AB}$ vs. $D^{al,\mathrm{N}_2}_{\mathrm{N}_2\rightarrow AB}$, and $D^{ver,\mathrm{N}_2}_{AB\rightarrow \mathrm{N}_2}$ vs. $D^{al,\mathrm{N}_2}_{AB\rightarrow \mathrm{N}_2}$ are presented. The performance of the alchemical derivatives can be significantly improved by increasing the ζ- splitting and adding the diffuse functions. However, the most crucial issue for the quantitative description of this type of transmutations turns out to be the presence of the tight functions in the basis set. The change in the nuclear screening and the core electron interaction is better predicted by the alchemical derivative if this augmentation is used. In case of the $N_2 \rightarrow CO$ transmutation, the error is around one mHartree (Table IV-3) illustrating that in this more complicated alchemical transmutation as compared to the deprotonation in Sec.IV.A, now involving two non H-atom of importance when exploring eg. benzene rings (Sec.IV.C), the transmutation energy, and certainly its trend, can very well be obtained using alchemical derivatives thus offering additional possibilities when exploring CS. The results obtained with the polarization consistent basis sets are worse than in case of the correlation consistent basis.

Table IV-3. The vertical and alchemical transmutation energies and errors (Δ) between the alchemical prediction and vertical calculation for the $N_2 \rightleftarrows CO$ transmutation type. All data in a.u.

		transmutation type: $N_2 \rightleftarrows CO$				
		$D^{\text{ver},N_2}_{N_2\to CO}$	$D^{\text{al},N_2}_{N_2\to CO}$	$D^{\text{al},N_2}_{CO\to N_2}$	$\Delta D^{\text{al},N_2}_{N_2\to CO}$	$\Delta D^{\text{al},N_2}_{CO\to N_2}$
B3LYP	cc-pVDZ	-3.785	-2.389	5.179	-1.396	1.394
	cc-pVTZ	-3.786	-3.462	4.151	-0.324	0.366
	aug -cc-pVDZ	-3.786	-2.861	4.672	-0.925	0.886
	aug -cc-pVTZ	-3.785	-3.605	3.956	-0.180	0.170
	cc-pCVDZ	-3.785	-3.666	3.919	-0.119	0.134
	cc-pCVTZ	-3.786	-3.782	3.802	**-0.004**	**0.017**
	aug -cc-pCVDZ	-3.786	-3.683	3.899	-0.102	0.114
	aug -cc-pCVTZ	-3.785	-3.784	3.801	**-0.001**	**0.016**
PBE	cc-pVDZ	-3.778	-2.386	5.161	-1.392	1.383
	cc-pVTZ	-3.779	-3.450	4.143	-0.329	0.364
	aug -cc-pVDZ	-3.778	-2.858	4.659	-0.921	0.881
	aug -cc-pVTZ	-3.779	-3.593	3.949	-0.185	0.170
	cc-pCVDZ	-3.778	-3.658	3.908	-0.119	0.130
	cc-pCVTZ	-3.779	-3.775	3.793	**-0.004**	**0.014**
	aug -cc-pCVDZ	-3.778	-3.676	3.889	-0.102	0.111
	aug -cc-pCVTZ	-3.779	-3.777	3.792	**-0.002**	**0.013**

IV.C Benzene's transmutation

Using benzene as the reference molecule, two types of transmutations are studied. The first one is the substitution of the N units for (C,H) units which leads to azines at the benzene geometry,

$$\left(\mathrm{C{,}H}\right)_n \to \mathrm{N}_n \,. \tag{4-12}$$

The second one is the substitution of (C,C) units by the isoelectronic (B,N) units which leads to the azaborines:

$$\left(\mathrm{C{,}C}\right)_n \to \left(\mathrm{B{,}N}\right)_n \,. \tag{4-13}$$

Azines are important because of their use as powerful reagents in synthetic organic chemistry, where their aromaticity is at stake. Azaborines are also important in synthetic organic chemistry, but recently the incorporation of boron as a part of a functional target structure has emerged as a useful way to generate diversity in organic compounds. The replacement of a (C,C) unit in an aromatic molecule with an isoelectronic (B,N) unit has been shown to impart interesting electronic, photophysical, luminescent, and chemical properties, which are often very distinct from those of the (C,C) containing aromatic analogues.[104-106] The syntheses of the azaborine compounds are, in general, very challenging and often involve multi-step reactions or the use of the transition-metal catalysts.[107,108] Thus, prediction of the most stable (B,N) -containing compounds can be very useful for the development of efficient and simple synthetic methods of this important class of compounds.

In this part, the same basis sets as described above were used. As the exchange-correlation functional only B3LYP functional was used which is known as the most widely used functional in trouble-free situations (covalent interaction, excluding π-π stacking and no transition metal bonds) and which proved to yield highly satisfactory results in IV.A and IV.B. In addition, the calculations at HF/STO-3 level were included. As it can be expected, the HF method predictions are quantitatively wrong, but the main point of these calculations is their qualitative verification as preliminary results for complex CS scanning.

IV.C.1.a Azines.

In general, the energy change for the transmutation which is a combination of the proton annihilation and the replacement of a carbon by a nitrogen atom in benzene is

$$D^{\mathrm{al},i}_{\mathrm{C_6H_6}} = \mu^{\mathrm{al}}_{\mathrm{H}} - \mu^{\mathrm{al}}_{\mathrm{C}} + \frac{1}{2}\left(\eta^{\mathrm{al}}_{\mathrm{C,C}} + \eta^{\mathrm{al}}_{\mathrm{H,H}}\right) - \eta^{\mathrm{al}}_{\mathrm{C_1,H}_i} \,. \tag{4-14}$$

Due to the benzene symmetry, the first derivatives are equal for the given type of atom. The same is true for the diagonal terms of the second derivative matrix. Assuming that the charge changes on the carbon atom numbered as one, the most stable product of the transmutation is indicated by the highest value of the carbon-hydrogen alchemical hardness, $\eta^{al}_{\mathrm{C,H}}$. The highest value of $\eta^{al}_{\mathrm{C,H}}$ occurs for the hydrogen connected with the replaced carbon (in Table IV-4, the alchemical derivatives for the benzene at B3LYP/cc-pVDZ level are presented). Therefore, the most stable product of this transmutation is pyridine with transmutation energy:

$$D^{\mathrm{al}}_{(\mathrm{C-H})\to\mathrm{N}} = \mu^{al}_{\mathrm{H}} - \mu^{al}_{\mathrm{C}} + \frac{1}{2}\left(\eta^{al}_{\mathrm{C,C}} + \eta^{al}_{\mathrm{H,H}}\right) - \eta^{al}_{\mathrm{C1,H1}}\,. \tag{4-15}$$

The set of the transmutation vectors for the $\mathrm{C_6H_6} \to \mathrm{C}_{6\text{-}n}\mathrm{H}_{6\text{-}n}\mathrm{N}_n$ is

$$\begin{aligned}\mathcal{D}^{\mathrm{C}_{6\text{-}n}\mathrm{H}_{6\text{-}n}\mathrm{N}_n}_{\mathrm{C_6H_6}} = \Big\{ d\mathbf{Z}^{12}_{\mathrm{C_6H_6}} = \left(dZ_{\mathrm{C1}},\ldots,dZ_{\mathrm{C6}},dZ_{\mathrm{H1}},\ldots,dZ_{\mathrm{H6}}\right) \Big| \, dZ_{\mathrm{C}i} \in \{0,1\}, dZ_{\mathrm{H}i} \in \{0,-1\}, \\ \sum_{i=1}^{6} dZ_{\mathrm{C}i} = -\sum_{i=1}^{6} dZ_{\mathrm{H}i} = n \Big\}\end{aligned} \tag{4-16}$$

and the transmutation energy is

$$\begin{aligned}D^{\mathrm{C}_{6\text{-}n}\mathrm{H}_{6\text{-}n}\mathrm{N}_n}_{\mathrm{C_6H_6}}\left[d\mathbf{Z}\right] = n\left(\mu^{al}_{\mathrm{H}} - \mu^{al}_{\mathrm{C}} + \frac{1}{2}\left(\eta^{al}_{\mathrm{C,C}} + \eta^{al}_{\mathrm{H,H}}\right)\right) \\ + \sum_{i=1}^{6}\sum_{j>i}^{6}\left(\eta^{al}_{\mathrm{C}i,\mathrm{C}j} dZ_{\mathrm{C}i} dZ_{\mathrm{C}j} + \eta^{al}_{\mathrm{H}i,\mathrm{H}j} dZ_{\mathrm{H}i} dZ_{\mathrm{H}j}\right) + \sum_{i=1}^{6}\sum_{j=1}^{6} \eta^{al}_{\mathrm{H}i,\mathrm{C}j} dZ_{\mathrm{C}i} dZ_{\mathrm{H}j}\,.\end{aligned} \tag{4-17}$$

The first term in Eq.(4-17) is constant for given *n*, the second term brings a positive contribution to the sum, while the last term is negative. From simple examination of the carbon-hydrogen hardness values ($\eta^{al}_{\mathrm{H}i,\mathrm{C}j}$ elements in Table IV-4) for the given carbon replacements, $d\mathbf{Z}_{\mathrm{C}} = \left(dZ_{\mathrm{C1}},\ldots,dZ_{\mathrm{C6}}\right)$, the annihilation of the proton connected with these carbons minimizes the last term in Eq.(4-17), $\eta^{\mathrm{al}}_{\mathrm{H}i,\mathrm{C}i} = 0.338$. This means that $\left(dZ_{\mathrm{H1}},\ldots,dZ_{\mathrm{H6}}\right) = -\left(dZ_{\mathrm{C1}},\ldots,dZ_{\mathrm{C6}}\right)$, and Eq.(4-17) can be rewritten as

$$\begin{aligned}D^{\mathrm{al}}_{(\mathrm{C-H})_n\to\mathrm{N}_n}\left[d\mathbf{Z}_{\mathrm{C}}\right] \equiv D^{\mathrm{C}_{6\text{-}n}\mathrm{H}_{6\text{-}n}\mathrm{N}_n}_{\mathrm{C_6H_6}}\left[d\mathbf{Z} = \left(d\mathbf{Z}_{\mathrm{C}}, -d\mathbf{Z}_{\mathrm{C}}\right)\right] \\ = nD^{\mathrm{al}}_{(\mathrm{C-H})\to\mathrm{N}} + \sum_{i=1}^{6}\sum_{j>i}^{6}\left(\eta^{al}_{\mathrm{C}i,\mathrm{C}j} + \eta^{al}_{\mathrm{H}i,\mathrm{H}j} - 2\eta^{al}_{\mathrm{C}i,\mathrm{H}j}\right) dZ_{\mathrm{C}i} dZ_{\mathrm{C}j}\,.\end{aligned} \tag{4-18}$$

The change in the interaction energy between two C-H units after their transmutation is $\left(\eta^{al}_{\mathrm{C}i,\mathrm{C}j} + \eta^{al}_{\mathrm{H}i,\mathrm{H}j} - 2\eta^{al}_{\mathrm{C}i,\mathrm{H}j}\right)$. Using data from Table IV-4, this change is 0.035, -0.002, 0.004 a.u., when $\mathrm{C}j$ is at the ortho, meta and para position with respect to $\mathrm{C}i$, respectively. This result means that to obtain the most stable azines, carbons should be replaced in such a way that maximizes the number of carbon pairs at meta position with respect to themselves and minimizes the number of carbon pairs at ortho position with respect to themselves. In Table IV-6 we present the transmutation energies for the $(\mathrm{C-H})_n \to \mathrm{N}_n$ transmutations (and their components) which confirm the above statement. In the case of the diazines, triazines and tetrazines there are three isomers. In all cases, the structure with the maximum number of carbon pairs at meta position with respect to themselves is the most stable for the given *n*: the 1,3-diazine (one pair at (1,3) position), 1,3,5-triazine (three pairs at (1,3),(1,5),(3,5) positions) and 1,2,3,5-tetrazine. The 1,2,4-triazine is more stable than 1,2,3-triazine - both have one pair of non C-atoms at meta-position with respect to themselves, but the former has one pair at ortho position with respect to themselves while the latter has two. The nuclear components of D^{al}, which depend only on the geometry and the nuclear charges ($D^{\mathrm{al,nuc}}, D^{\mathrm{al},\mu,\mathrm{nuc}}, D^{\mathrm{al},\eta,\mathrm{nuc}}$ from Eq.(3-32)), can be used as preliminary indicators. However, their predictions are not always consistent with the D^{al} ordering e.g. the $D^{\mathrm{al,nuc}}$ ordering is consistent with the D^{al} ordering for the triazines, but inconsistent in the case of the diazines and tetrazines.

The MAE between D^{al} and $D^{\mathrm{ver}}, D^{\mathrm{ver,BSSE}}$, D^{rel} are collected in Table IV-5. Inclusion of the ghost atom at the deprotonation site has very little effect on the MAE. In some cases, we observe smaller MAE values between D^{al} and D^{rel} than between D^{al} and D^{ver}. This can be explained as an effect of the cancellation of errors connected with the higher-order terms and the geometry relaxation. In general, however, the basis set dependence is complex. The lowest MAE is observed for the cc-pVTZ basis and the introduction of the tight functions significantly increases the MAE. Increase of the ζ- splitting without the tight functions gives the opposite result. The effect of the diffuse functions seems to be inexplicable at present.

The analysis of this transmutation can illuminate these results. The $(\mathrm{C-H})_n \to \mathrm{N}_n$ transmutation can be split into two steps: the carbon atom to nitrogen cation transmutation and the deprotonation as follows

$$\begin{tikzcd}
(\mathrm{C-H})_n \arrow[r] \arrow[d] \arrow[dr] & (\mathrm{N-H})_n^{n+} \arrow[d] \\
\mathrm{C}_n^{n-} \arrow[r] & \mathrm{N}_n
\end{tikzcd} \tag{4-19}$$

Since problems connected with the highly negative ion calculations can be expected, only for the $(\mathrm{C-H})_n \to (\mathrm{N-H})_n^{n+}$ transmutation, the vertical transmutation energies were calculated. The MAEs are presented in Table IV-5. These results are in line with the results obtained for the $(\mathrm{C,C}) \to (\mathrm{B,N})$ transmutation and the nitrogen molecule transmutation, which suggests that the basis set dependency for the complex transmutation is not simply predictable from the basis set dependencies of its components. Alchemical derivatives should be used for qualitative rather than quantitative predictions in these types of transmutations which are essential in comparative studies exploring CS. The quantitative prediction requires special attention to the basis set selection.

Table IV-4. Alchemical derivatives and theirs electronic and nuclear components for benzene at B3LYP/cc-pVDZ level. All values in a.u.

atom	first derivatives		
	$\mu^{\mathrm{al,el}}$	$\mu^{\mathrm{al,nuc}}$	μ^{al}
C	-24.453	9.713	-14.740
H	-10.388	9.291	-1.097
bond	second derivatives		
	$\eta^{\mathrm{al,el}}$	$\eta^{\mathrm{al,nuc}}$	η^{al}
C1-C1	-1.900	0.000	-1.900
C1-C2	-0.102	0.378	0.276
C1-C3	-0.006	0.218	0.213
C1-C4	0.029	0.189	0.219
C1-H1	-0.146	0.484	0.338
C1-H2	-0.013	0.245	0.231
C1-H3	0.048	0.155	0.203
C1-H4	0.064	0.136	0.200
H1-H1	-1.087	0.000	-1.087
H1-H3	0.008	0.212	0.221
H1-H3	0.068	0.123	0.191
H1-H4	0.079	0.106	0.185

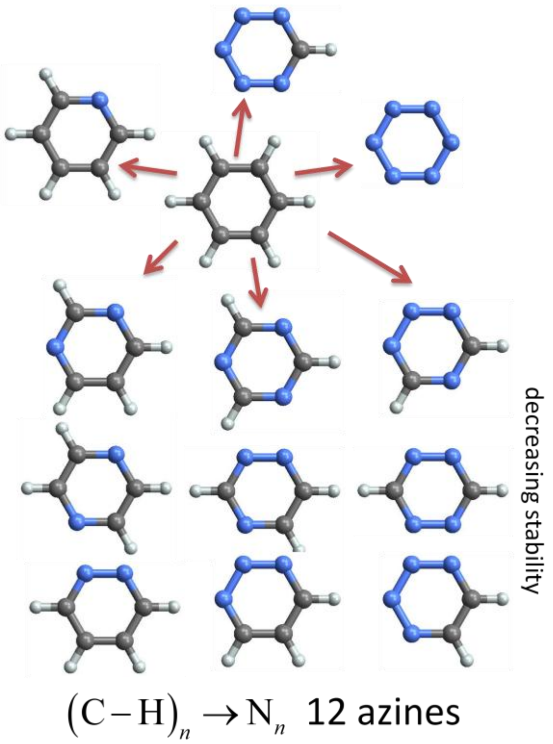

Fig. IV-2 Stable isomers of azines of the $(\mathrm{C-H})_n \rightarrow \mathrm{N}_n$ quasi-type transmutation for benzene at B3LYP/cc-pVDZ level.

IV.C.1.b Azaborines.

In case of the $(\mathrm{C,C}) \rightarrow (\mathrm{B,N})$ transmutation, the odd terms in the Taylor expansion cancel as in case of the $\mathrm{N_2} \rightarrow \mathrm{CO}$ transmutation and the replacement energy of two carbons by one nitrogen and one boron is simply given by

$$D^{\mathrm{al},i}_{\mathrm{C_6H_6}} = \eta^{\mathrm{al}}_{\mathrm{C,C}} - \eta^{\mathrm{al}}_{\mathrm{C_1,C}_i} , \qquad (4\text{-}20)$$

and the most preferable substitution is to replace two adjacent carbon (see Table IV-4). The set of the transmutation vectors for the $C_6H_6 \to C_{6\text{-}2n}H_6(BN)_n$ transmutation is

$$\mathcal{D}_{C_6H_6}^{C_{6\text{-}2n}H_6(BN)_n} = \left\{ d\mathbf{Z}_{C_6H_6}^{12} = (dZ_{C1},...,dZ_{C6},0,...,0) \middle| dZ_{Ci} \in \{-1,0,1\}, \sum_{i=1}^{6} dZ_{Ci} = 0 \right\} \quad (4\text{-}21)$$

and the transmutation energy is

$$\begin{aligned} D_{(C,C)\to(B,N)}^{al}[d\mathbf{Z}_C] &\equiv D_{C_6H_6}^{C_{6-2n}H_6(BN)_n}\left[d\mathbf{Z} = (d\mathbf{Z}_C,0)\right] \\ &= n\eta_{C,C}^{al} + \sum_{i=1}^{6}\sum_{j=1}^{6}\eta_{Ci,Cj}^{al} dZ_{Ci} dZ_{Cj} \end{aligned} . \quad (4\text{-}22)$$

To simplify the analysis of the above equation, the $\mathbf{Z}_C$ will be replaced by two vectors

$$\mathbf{N}^n = \left\{ (N1,...,Nn) \middle| Ni < Nj, Z_{Ni} = 1 \right\}, \quad \mathbf{B}^n = \left\{ (B1,...,Bn) \middle| Bi < Bj, Z_{Bi} = 1 \right\}, \quad (4\text{-}23)$$

and Eq.(4-22) can be rewritten as

$$D^{al}\left[\mathbf{N}^n, \mathbf{B}^n\right] = n\,\eta_{C,C}^{al} + \sum_{i=1}^{n}\sum_{j>i}^{n}\left(\eta_{Ni,Nj}^{al} + \eta_{Bi,Bj}^{al}\right) - \sum_{i=1}^{n}\sum_{j=1}^{n}\eta_{Ni,Bj}^{al} . \quad (4\text{-}24)$$

Let us relabel the carbon-carbon alchemical hardness as follows: $\eta_{ortho}^{al} = \eta_{C1,C2}^{al}$, $\eta_{meta}^{al} = \eta_{C1,C3}^{al}$, $\eta_{para}^{al} = \eta_{C1,C4}^{al}$, and note that $\eta_{ortho}^{al} > \eta_{para}^{al} > \eta_{meta}^{al}$. The preferable position for the pair of heteroatoms is the ortho-position with respect to themselves followed by para and meta. Based on this conclusion the results presented in Table IV-7 for the azaborines become easy to understand. In case $n = 1$, the 1,2-azaborine is the most stable, the 1,4-azaborine with the substitution at the para -position is more stable than the 1,3-azaborine (meta -position). In case $n = 2$, the last term in Eq.(4-24) is a combination of six components, the second is a sum of the nitrogen-nitrogen alchemical hardness and the boron-boron alchemical hardness. The substitution of the next four carbons causes the appearance of a combination of three η_{ortho}^{al}, one η_{para}^{al} and two η_{meta}^{al} in Eq.(4-24). The $\mathbf{N}^2 = (1,3)$ and $\mathbf{B}^2 = (2,4)$ vectors are a minimizer of Eq.(4-22), $D^{al}\left[(1,3),(2,4)\right] = 2\eta_{C,C}^{al} + 2\eta_{meta}^{al} - \left(3\eta_{ortho}^{al} + \eta_{para}^{al}\right)$.

Table IV-5. Mean absolute error between $D^{\mathrm{al}}_{\mathrm{C_6H_6}}$ and the vertical transmutation energy, $D^{\mathrm{ver}}_{\mathrm{C_6H_6}}$, the vertical transmutation energy with the ghost function at deprotonation site, $D^{\mathrm{ver,BSSE}}_{\mathrm{C_6H_6}}$, and the transmutation energy (including geometry relaxation), $D^{\mathrm{rel}}_{\mathrm{C_6H_6}}$. All data in a.u.

		Transmutation type					
		$(\mathrm{C-H})_n \to \mathrm{N}_n$			$\mathrm{C}_n \to \mathrm{N}_n$	$(\mathrm{C,C})_n \to (\mathrm{B,N})_n$	
		$D^{\mathrm{ver}}_{\mathrm{C_6H_6}}$	$D^{\mathrm{ver,BSSE}}_{\mathrm{C_6H_6}}$	$D^{\mathrm{rel}}_{\mathrm{C_6H_6}}$	$D^{\mathrm{ver}}_{\mathrm{C_6H_6}}$	$D^{\mathrm{ver}}_{\mathrm{C_6H_6}}$	$D^{\mathrm{rel}}_{\mathrm{C_6H_6}}$
B3LYP	cc-pVDZ	1.802	1.809	1.825	2.221	2.466	2.532
	cc-pVTZ	**0.039**	**0.036**	**0.016**	0.414	0.355	0.419
	aug -cc-pVDZ	0.920	0.921	0.944	1.329	1.431	1.495
	aug -cc-pVTZ	0.309	0.309	0.287	0.128	0.052	0.116
	cc-pCVDZ	0.089	0.082	0.065	0.325	0.203	0.268
	cc-pCVTZ	0.346	0.344	0.323	**0.108**	**0.017**	**0.075**
	aug -cc-pCVDZ	0.128	0.127	0.104	0.280	0.184	0.247
	aug -cc-pCVTZ	0.349	0.349	0.327	**0.101**	**0.016**	**0.074**
HF	STO-3	2.817	2.895	2.858	3.714	3.701	3.768

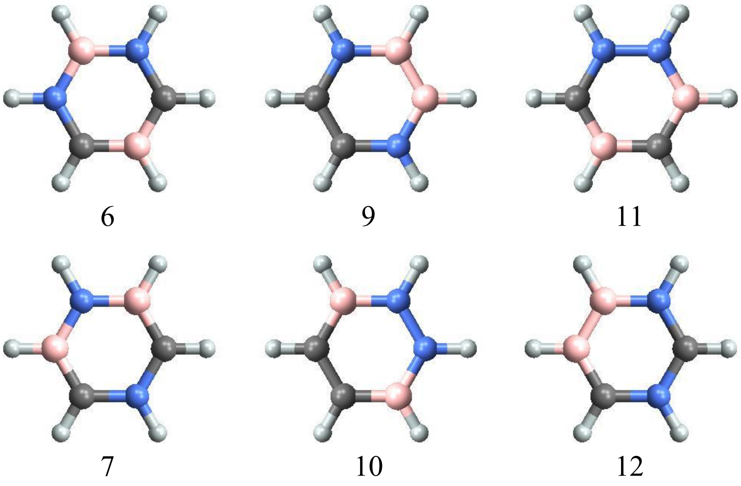

Fig. IV-3. The azaborine pairs (6,7), (9,10), (11,12) for which the alchemical transmutation energy is equal (see Table IV-7)

Table IV-6. Transmutation energy for $\left(\mathrm{C}-\mathrm{H}\right)_n \rightarrow \mathrm{N}_n$ transmutation type at B3LYP/ cc-pVTZ level. $D_{\mathrm{C_6H_6}}^{\mathrm{al}}$ - alchemical prediction for transmutation energy; and its components (see Eq.(3-32)). $\Delta D_{\mathrm{C_6H_6}}^{\mathrm{ver}}$, $\Delta D_{\mathrm{C_6H_6}}^{\mathrm{ver,BSSE}}$, $\Delta D_{\mathrm{C_6H_6}}^{\mathrm{rel}}$ - difference between the alchemical transmutation energy and the vertical transmutation energy, the vertical transmutation energy with the ghost function at deprotonation site as well as the difference between the transmutation energy and geometry relaxation, respectively. All data in a.u.

n	name	$D_{\mathrm{C_6H_6}}^{\mathrm{al}}$	$D_{\mathrm{C_6H_6}}^{\mathrm{al},\mu}$	$D_{\mathrm{C_6H_6}}^{\mathrm{al},\eta}$	$D_{\mathrm{C_6H_6}}^{\mathrm{al,el}}$	$D_{\mathrm{C_6H_6}}^{\mathrm{al},\mu,\mathrm{el}}$	$D_{\mathrm{C_6H_6}}^{\mathrm{al},\eta,\mathrm{el}}$	$D_{\mathrm{C_6H_6}}^{\mathrm{al,nuc}}$	$D_{\mathrm{C_6H_6}}^{\mathrm{al},\mu,\mathrm{nuc}}$	$D_{\mathrm{C_6H_6}}^{\mathrm{al},\eta,\mathrm{nuc}}$	$\Delta D_{\mathrm{C_6H_6}}^{\mathrm{ver}}$	$\Delta D_{\mathrm{C_6H_6}}^{\mathrm{ver,BSSE}}$	$\Delta D_{\mathrm{C_6H_6}}^{\mathrm{rel}}$
1	pyridine	-16.043	-13.652	-2.390	-15.959	-14.058	-1.902	-0.084	0.405	-0.489	0.017	0.016	0.008
2	1,3-diazine	-32.089	-27.305	-4.784	-31.952	-28.115	-3.837	-0.136	0.811	-0.947	0.031	0.029	0.011
	1,4-diazine	-32.082	-27.305	-4.778	-31.939	-28.115	-3.823	-0.144	0.811	-0.954	0.029	0.027	0.011
	1,2-diazine	-32.054	-27.305	-4.750	-31.989	-28.115	-3.873	-0.066	0.811	-0.876	0.025	0.024	0.012
3	1,3,5-triazine	-48.137	-40.957	-7.180	-47.979	-42.173	-5.806	-0.158	1.216	-1.374	0.047	0.045	0.015
	1,2,4-triazine	-48.097	-40.957	-7.140	-48.002	-42.173	-5.829	-0.095	1.216	-1.311	0.040	0.038	0.016
	1,2,3-triazine	-48.069	-40.957	-7.112	-48.052	-42.173	-5.879	-0.017	1.216	-1.233	0.031	0.028	0.014
4	1,2,3,5-tetrazine	-64.114	-54.609	-9.505	-64.099	-56.231	-7.868	-0.015	1.622	-1.637	0.049	0.046	0.019
	1,2,4,5-tetrazine	-64.108	-54.609	-9.498	-64.085	-56.231	-7.854	-0.022	1.622	-1.644	0.054	0.051	0.023
	1,2,3,4-tetrazine	-64.080	-54.609	-9.471	-64.135	-56.231	-7.905	0.056	1.622	-1.566	0.038	0.036	0.017
5	pentazine	-80.094	-68.262	-11.832	-80.253	-70.289	-9.964	0.159	2.027	-1.868	0.050	0.047	0.022
6	hexazine	-96.077	-81.914	-14.163	-96.440	-84.346	-12.094	0.364	2.432	-2.069	0.052	0.049	0.024

Table IV-7. Transmutation energy for $(\mathrm{C},\mathrm{C})\rightarrow(\mathrm{B},\mathrm{N})$ transmutation type at B3LYP/ cc-pCVTZ level $D^{\mathrm{al}}_{\mathrm{C_6H_6}}$ - alchemical prediction for transmutation energy; and its components (see Eq.(3-32)). $D_{\mathrm{N}}^{\mathrm{al},\eta,\mathrm{nuc}}$ and $D_{\mathrm{B}}^{\mathrm{al},\eta,\mathrm{nuc}}$ - the homoatom parts of the $D^{\mathrm{al},\eta,\mathrm{nuc}}$ (see Eq. (4-24)). $\Delta D^{\mathrm{ver}}_{\mathrm{C_6H_6}}$, $\Delta D^{\mathrm{rel}}_{\mathrm{C_6H_6}}$ - difference between the alchemical transmutation energy and the vertical transmutation energy, the transmutation energy and geometry relaxation, respectively. All data in a.u.

n		name	$D^{\mathrm{al}}_{\mathrm{C_6H_6}}$	$D^{\mathrm{al,el}}_{\mathrm{C_6H_6}}$	$D^{\mathrm{al,nuc}}_{\mathrm{C_6H_6}}$	$D_{\mathrm{N}}^{\mathrm{al},\eta,\mathrm{nuc}}$	$D_{\mathrm{B}}^{\mathrm{al},\eta,\mathrm{nuc}}$	$\Delta D^{\mathrm{ver}}_{\mathrm{C_6H_6}}$	$\Delta D^{\mathrm{rel}}_{\mathrm{C_6H_6}}$
1	1	1,2-azaborine	-3.406	-3.025	-0.380	0.000	0.000	-0.001	-0.027
	2	1,4-azaborine	-3.354	-3.164	-0.190	0.000	0.000	-0.002	-0.044
	3	1,3-azaborine	-3.347	-3.128	-0.220	0.000	0.000	-0.002	-0.038
2	4	1,3,2,4-diazadiborine	-6.877	-5.984	-0.892	0.220	0.220	0.003	-0.038
	5	1,5,2,4- diazadiborine	-6.825	-6.123	-0.702	0.220	0.220	0.024	-0.036
	6	1,3,2,5- diazadiborine	-6.805	-6.015	-0.790	0.220	0.190	-0.005	-0.049
	7	1,4,2,6- diazadiborine	-6.805	-6.015	-0.790	0.190	0.220	-0.001	-0.052
	8	1,4,2,5- diazadiborine	-6.799	-5.979	-0.820	0.190	0.190	-0.005	-0.051
	9	1,4,2,3- diazadiborine	-6.747	-6.117	-0.629	0.380	0.190	-0.034	-0.087
	10	1,2,3,6- diazadiborine	-6.747	-6.117	-0.629	0.190	0.380	-0.025	-0.099
	11	1,2,3,5- diazadiborine	-6.701	-6.291	-0.410	0.220	0.380	-0.005	-0.092
	12	1,3,4,5- diazadiborine	-6.701	-6.291	-0.410	0.380	0.220	-0.012	-0.078
	13	1,2,3,4- diazadiborine	-6.643	-6.394	-0.249	0.380	0.380	-0.037	-0.126
	14	1,2,4,5- diazadiborine	-6.591	-6.532	-0.059	0.380	0.380	-0.021	-0.119
3	15	1,3,5,2,4,6-triazatriborinane	-10.412	-8.877	-1.535	0.659	0.659	0.019	-0.023
	16	1,2,5,3,4,6-triazatriborinane	-10.152	-9.143	-1.010	0.790	0.790	-0.050	-0.139
	17	1,2,3,4,5,6-triazatriborinane	-9.944	-9.695	-0.249	0.980	0.980	-0.040	-0.181

The 1,4,2,3-diazadiborine and the 1,2,3,6-diazadiborine are less stable than the 1,3,2,4-diazadiborine. The alchemical transmutation energy for the former two diazaborines is $D^{al}\left[(1,4),(2,3)\right]=D^{al}\left[(1,2),(3,6)\right]=2\eta^{\mathrm{al}}_{\mathrm{C,C}}+\eta^{\mathrm{al}}_{\mathrm{para}}+\eta^{\mathrm{al}}_{\mathrm{ortho}}-\left(2\eta^{\mathrm{al}}_{\mathrm{ortho}}+2\eta^{\mathrm{al}}_{\mathrm{meta}}\right)$. This difference in the transmutation energy is an effect of the change from the $N-B-N-B$ to the $B-N-N-B$ or $N-B-B-N$ sequence. Both of these diazaborines have the same value of D^{al} since the second-order Taylor expansion cannot distinguish between these two molecules. The same situation occurs in the case of the 1,3,2,5-diazaborine and 1,4,2,6-diazaborine and in the case of the 1,2,3,5-diazaborine and 1,3,4,5-diazaborine. The splitting into the nitrogen-nitrogen interaction and the boron-boron interaction terms:

$$D^{\mathrm{al},\eta,\mathrm{nuc}}_{\mathrm{B}}\left[\mathbf{N}^n,\mathbf{B}^n\right]=\sum_{i=1}^{n}\sum_{j>i}^{n}\eta^{\mathrm{al}}_{\mathrm{B}i,\mathrm{B}j},\qquad D^{\mathrm{al},\eta,\mathrm{nuc}}_{\mathrm{N}}\left[\mathbf{N}^n,\mathbf{B}^n\right]=\sum_{i=1}^{n}\sum_{j>i}^{n}\eta^{\mathrm{al}}_{\mathrm{N}i,\mathrm{N}j}, \qquad (4\text{-}25)$$

can be used to distinguish these cases. In Table IV-7, the splitting of the nuclear components, which involve only geometrical data, was used to distinguish isomers presented in Fig. IV-3.

In case of azaborines, the basis set dependence is very easy to figure out (see Table IV-5. Mean absolute error between $D^{\mathrm{al}}_{\mathrm{C_6H_6}}$ and the vertical transmutation energy, $D^{\mathrm{ver}}_{\mathrm{C_6H_6}}$, the vertical transmutation energy with the ghost function at deprotonation site, $D^{\mathrm{ver,BSSE}}_{\mathrm{C_6H_6}}$, and the transmutation energy (including geometry relaxation), $D^{\mathrm{rel}}_{\mathrm{C_6H_6}}$. All data in a.u. Table IV-5). Increasing the ζ- splitting results in a decrease of the MAE. As in the case of the nitrogen molecule, the introduction of the tight functions in the basis set is crucial for the quantitative description. The MAEs for the cc-pCVTZ basis and its augmented counterpart are very similar, so that cc-pCVTZ basis is advised, especially when one takes into account the average calculation timings. Of course, the HF/STO-3 results should not be used for a quantitative prediction, but as we can see from Fig. IV-4, they are in qualitative agreement with the B3LYP/cc-pCVTZ results. This is caused by the fact that the difference between the benzene geometry calculated at these two levels is very small. The correlation coefficient between the $D^{\mathrm{al,nuc}}$ at these two levels is 1, the intercept equals to zero, and their MEA is 0.006 in a.u. In addition, the effect of the difference in the methods used for the geometry optimization and the alchemical derivatives calculation was investigated for all levels listed in

Table IV-5. The MAE between D^{al} calculated at the same level but at different geometry level is less than 0.004 a.u. (one exception is the MAE between D^{al} from cc-pVTZ//cc-pVTZ and D^{al} cc-pVDZ//cc-pVTZ calculation, which is equal to 0.012 a.u.). General conclusion from these calculations is that geometry optimizations can be done at a less computationally demanding level.

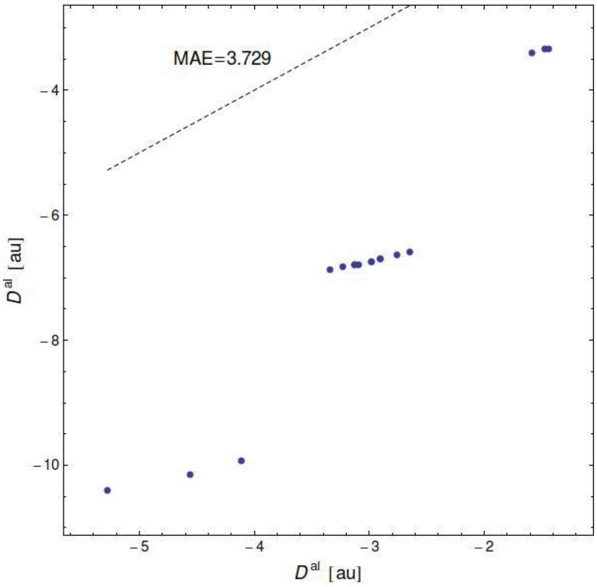

Fig. IV-4 Relation between the alchemical transmutation energy HF/STO-3 level (horizontal axis) vs. the alchemical transmutation energy at B3LYP/ cc-pVTZ level (vertical axis) for $(\mathrm{C,C}) \to (\mathrm{B,N})$ transmutation type. These two levels are qualitatively related

IV.D Pyrene case.

Pyrene is a ubiquitous fluorophore that has found application in various sensors and as a probe in biochemical labeling studies. B-N substitution provides new opportunities for organic electronics.[108] Examples of synthesized BN-pyrenes are very scarce, only 4,10,5,9-diazadiborapyrene,[109] 4,5-azaborapyrene[110] and 10b,10c-azaborapyrene[111] are known as yet.

In this section, the usefulness of the alchemical derivatives will be tested on the pyrene transmutation to the azabora-pyrenes: $\mathrm{C_{16}H_{10}} \to \mathrm{C_{16\text{-}2\mathit{n}}H_{10}(BN)}_n$. Pyrene is a tetracyclic aromatic hydrocarbon, the smallest polycyclic aromatic hydrocarbon (PAH) on which the effect of a various boron and nitrogen substituents and their positions (internal/periphery) in the PAH can be studied. Pyrene with a molecular formula of $C_{16}H_{10}$ and D_{2h} symmetry has five no-symmetry equivalent carbons (see Fig. IV-5). The alchemical energy of the transmutation of two carbons into one nitrogen and one boron is

$$D^{\mathrm{al},i,j}_{\mathrm{C_{16}H_{10}}} = \left(\mu_{\mathrm{C}i} - \mu_{\mathrm{C}j}\right) + \frac{1}{2}\left(\eta^{\mathrm{al}}_{\mathrm{C}i,\mathrm{C}i} + \eta^{\mathrm{al}}_{\mathrm{C}j,\mathrm{C}j} - 2\eta^{\mathrm{al}}_{\mathrm{C}i,\mathrm{C}j}\right). \qquad (4\text{-}26)$$

with an assumption that the nitrogen is placed at the position i and the boron at the position j. If substituted carbons are symmetry equivalent, then the contribution from the first derivative

vanishes and $D^{\mathrm{al},i,j} = D^{\mathrm{al},j,i}$. Otherwise, the preferable position for boron is the site with the highest value of the alchemical potential, ($\mu_{\mathrm{C}i} < \mu_{\mathrm{C}}$), e.g. $\mu_3 < \mu_{3\mathrm{a}}$ and the 3,3a-azaborapyrene (see
Fig. IV-5 and Fig. IV-7, **12**) is more stable than 3a,3-azaborapyrene. In case of 4,10,5,9-diazdiborapyrene, all substituted carbons are symmetry equivalent, so that the transmutation energy is

$$D^{\mathrm{al}}\left[(4,10),(5,9)\right] = 2\left(\eta^{\mathrm{al}}_{4,4} - \eta^{\mathrm{al}}_{4,5}\right) + 2\left(\eta^{\mathrm{al}}_{4,10} - \eta^{\mathrm{al}}_{4,9}\right), \qquad (4\text{-}27)$$

where $\left(\eta^{\mathrm{al}}_{4,4} - \eta^{\mathrm{al}}_{4,5}\right)$ is the $\left(\mathrm{C,C}\right) \to \left(\mathrm{B,N}\right)$ transmutation energy at the positions (4,5) and (10,9), and $2\left(\eta^{\mathrm{al}}_{4,10} - \eta^{\mathrm{al}}_{4,9}\right)$ is the interaction part between these two moieties. This interaction part is equal to 0.0002 a.u. for the 4,10,5,9-diazadiborapyrene and -0.0002 a.u. for the 4,9,5,10-diazadiborapyrene at B3LYP/cc-pCVTZ level (the $\eta^{\mathrm{al}}_{4,10}$ and $\eta^{\mathrm{al}}_{4,9}$ are 0.1457 and 0.1456 in a.u., respectively). This means that the latter is more stable that the former. Interestingly, the less stable isomer was synthesized from 2,6-diaminobiphenyl, while the more stable isomer could not be obtained in this manner.[104]

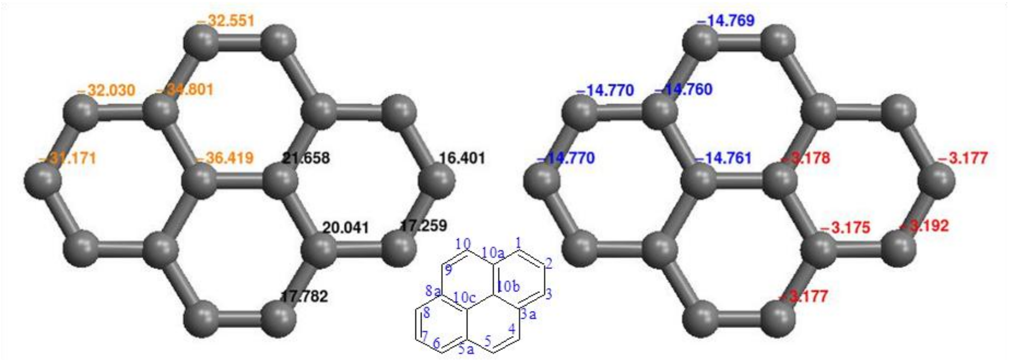

Fig. IV-5. Pyrene alchemical derivatives at B3LYP/ cc-pVTZ level. The alchemical potential (blue), the diagonal element of the alchemical hardness (red) and carbon label are on the left panel. The electronic (yellow) and nuclear (black) parts of the alchemical potential are on the right panel.

Based on the azaborine results, where the cc-pCVTZ results and the aug-cc-pCVTZ results are similar, the basis sets with augmented functions were not included in the test (see Table IV-5). As in the benzene case, the HF/STO-3 calculation was included to test its qualitative usefulness. The alchemical transmutation energy was calculated for all possible isomers (e.g. for n=1, there are 63 isomers). Then, for the 10 energetically most stable isomers, the calculations of the vertical transmutation energies were performed. As a measure of quantitative accuracy, the correlation coefficient and the MAE were used. The qualitative accuracy is characterized by three numbers: the compatibility number, the accuracy number and the deficiency number (see Appendix A for definitions). All these numbers are collected in Table IV-8. The smallest disagreement between the vertical and the alchemical predictions is observed at B3LYP/cc-pCVTZ level. The worst result (for n=7), after rejecting the most extreme case (the red point in Fig. IV-6), is in line with the other results. These results are consistent with those obtained for the transmutation of nitrogen and benzene: the double ζ-splitted basis sets are simply too small for this type of transmutation. Moreover, the presence of the tight functions in the basis set is crucial for the quantitative accuracy. For example, in case of the azaborapyrenes and diazadiborapyrenes, the MAE is only 2 mHartree. The HF results confirm their usefulness as a preliminary result (e.g. compare the deficiency numbers for the B3LYP/cc-pVDZ level and for the HF/STO-3 level).

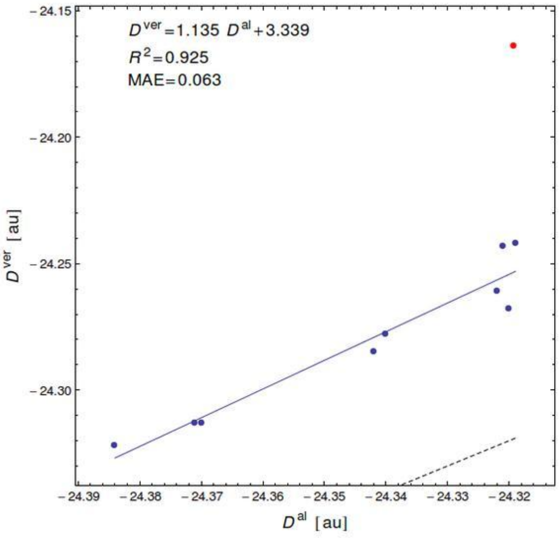

Fig. IV-6. Correlation between the alchemical transmutation energy vs. the vertical transmutation energy at B3LYP/ cc-pCVTZ level for $(\mathrm{C},\mathrm{C}) \rightarrow (\mathrm{B},\mathrm{N})$ transmutation type. The red point is not included.

The alchemical transmutation energies and their components are presented in Table IV-9 and the first three most stable isomers are presented in Fig. IV-7. In the case of the fullerenes, two major rules for BN-substitution were formulated: the “hexagon filling” and “continuity” rules. According to the former rule BN units tend to replace all three-carbon pairs of a hexagon one by one and then spread to adjacent hexagons. The “continuity” rule overshadows this rule where the incoming BN unit connects existing BN units to maintain the continuity of BN units. These two rules lead to an atomic arrangement where BN and C form separate regions in the network, consistent with several experimental investigations. This pattern of grouping heteroatoms and carbons is independent of the number of BN units in the fullerenes case. Beyond that, formation of unfavorable B-B and/or N-N bonds, or separation of BN units might destabilize the hybrid fullerene.[112] Inspection of the structures presented in Fig. IV-7, shows a similar but not identical pattern that, in most cases, individual hexagons are filled first followed by adjacent ones obeying a continuity rule, and that the avoidance of formation of unfavorable B-B and/or N-N bonds, decreasing the pyrene derivatives' stability, is confirmed. However, in the pyrene case, the most stable isomer is not always simply a combination of the incoming BN units with the existing BN groups.

The most stable diazadiborapyrene **21** is a combination of **12** and **13** azaborapyrenes. For *n*=3, the replacement of the interior carbon by the nitrogen is more preferable than by the boron (see ΔD_2^{al} for *n*=3 at Table IV-7 and **31**, **32** at Fig. IV-7). In the pyrene molecule, we can distinguish two types of rings: two rings with three fusion carbons (e.g. carbons at the positions 3a,10b and 10a) and two rings with four fusion carbons (e.g. carbons at the positions 3a,10b, 10c and 5a). These rings will be labeled *R3* and *R4*, respectively. In pyrene, at first the *R3* ring is filled and then the *R4* (**31** and **51**). The most stable isomer for *n*=6 suggests that in the next step the second *R4* ring will be filled but the most stable form for *n*=7 is the structure with two *R3* rings and one *R4* ring completely filled (compare **71** and **72**). The maximization of the nitrogen-boron connection rule is confirmed by isomer **81**. We should mention that these predictions are valid at the pyrene geometry and the geometry relaxation can change the observed ordering.

Table IV-8. Qualitative and quantitative accuracies of the alchemical prediction for the $(\mathrm{C},\mathrm{C})\rightarrow(\mathrm{B},\mathrm{N})$ transmutation type of pyrene. Item for given n and method contains: the correlation coefficient and the mean absolute error between D^{al} and the vertical transmutation energy, D^{ver} (upper part of item) and accuracy, compatibility, and deficiency numbers (lower part of item). The deficiency number is with respect to B3LYP/ cc-pCVTZ results.

		B3LYP								HF	
		cc-pVDZ		cc-pCVDZ		cc-pVTZ		cc-pCVTZ		STO-3	
n	1	0.107	1.194	0.989	0.098	0.879	0.153	**0.999**	**0.002**	0.988	1.806
		0–1–0		10–10–0		2–5–0		**10–10**		10–10–1	
	2	0.265	2.366	0.973	0.189	0.887	0.293	**0.982**	**0.002**	0.884	3.523
		0–0–4		3–6–0		0–2–3		**10–10**		3–4–1	
	3	0.765	3.542	0.987	0.275	0.861	0.437	**0.974**	**0.009**	0.887	5.203
		1–2–5		2–8–1		2–4–2		**2–8**		3–5–2	
	4	0.384	4.732	0.931	0.363	0.858	0.581	**0.868**	**0.015**	0.817	6.871
		1–3–3		3–4–1		3–4–3		**3–6**		4–4–6	
	5	0.851	5.913	0.918	0.441	0.845	0.717	**0.892**	**0.030**	0.807	8.515
		4–5–6		1–4–1		4–7–2		**4–5**		0–0–1	
	6	0.207	7.094	0.943	0.526	0.373	0.863	**0.937**	**0.038**	0.561	10.150
		1–1–3		1–4–0		3–5–2		**5–5**		5–7–3	
	7	0.483	8.265	0.597	0.589	0.490	0.978	**0.639**	**0.072**	0.426	11.749
		4–4–2		1–5–1		5–6-0		**5–5**		3–6–1	
	8	0.988	9.548	0.989	0.754	0.982	1.197	**0.988**	**0.022**	0.905	13.855
		2–8–0		2–4–0		2–7–2		**2–8**		2–4–1	
Total		1.000	5.332	1.000	0.404	1.000	0.652	**1.000**	**0.024**	0.998	7.709

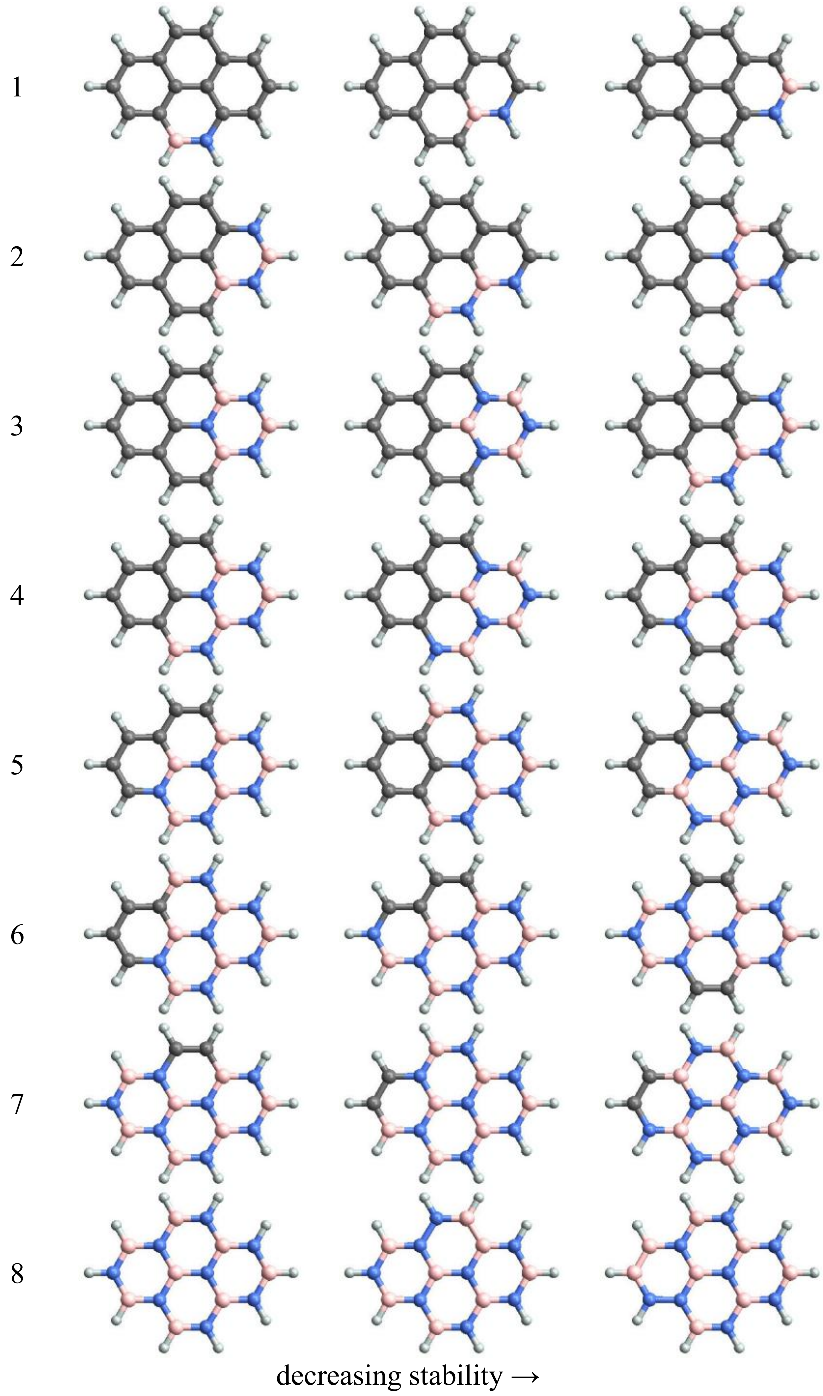

Fig. IV-7 The first three most stable isomers of the $(\mathrm{C},\mathrm{C}) \to (\mathrm{B},\mathrm{N})$ transmutation type of pyrene at B3LYP/ cc-pCVTZ level. The bold numbers used in text means the row and the item in this row, e.g. **12**, means the second isomer in the first row.

Table IV-9. The vertical transmutation energy, D^{ver}, the alchemical transmutation energy and its components for the most stable isomers of the $(\mathrm{C},\mathrm{C}) \to (\mathrm{B},\mathrm{N})$ transmutation type of pyrene at B3LYP/ cc-pCVTZ level. ΔD_2^{al} and ΔD_3^{al} are the difference between the most stable isomer and the second, and third isomers in the stability order (see Fig. IV-7)

	n							
	1	2	3	4	5	6	7	8
D^{ver}	-3.424	-6.886	-10.388	-13.858	-17.344	-20.814	-24.322	-27.834
D^{al}	-3.426	-6.891	-10.403	-13.885	-17.371	-20.856	-24.384	-27.922
$D^{\mathrm{al},\mu}$	0.000	-0.011	-0.012	-0.012	-0.011	-0.011	0.000	0.000
$D^{\mathrm{al},\eta}$	-3.426	-6.880	-10.391	-13.873	-17.360	-20.845	-24.384	-27.922
$D^{\mathrm{al,el}}$	-3.036	-4.076	-8.574	-11.585	-12.876	-15.908	-20.715	-23.708
$D^{\mathrm{al},\mu,\mathrm{el}}$	0.000	1.913	0.295	0.295	1.913	1.913	0.000	0.000
$D^{\mathrm{al},\eta,\mathrm{el}}$	-3.036	-5.989	-8.869	-11.880	-14.788	-17.821	-20.715	-23.708
$D^{\mathrm{al,nuc}}$	-0.390	-2.815	-1.828	-2.300	-4.496	-4.947	-3.670	-4.214
$D^{\mathrm{al},\mu,\mathrm{nuc}}$	0.000	-1.924	-0.307	-0.307	-1.924	-1.924	0.000	0.000
$D^{\mathrm{al},\eta,\mathrm{nuc}}$	-0.390	-0.891	-1.522	-1.993	-2.572	-3.024	-3.670	-4.214
$D_{\mathrm{N}}^{\mathrm{al},\eta,\mathrm{nuc}}$	0.000	0.219	0.655	1.208	1.873	2.658	3.494	4.607
$D_{\mathrm{B}}^{\mathrm{al},\eta,\mathrm{nuc}}$	0.000	0.218	0.652	1.103	1.873	2.647	3.494	4.607
ΔD_2^{al}	0.009	0.002	0.024	0.024	0.005	0.006	0.013	0.223
ΔD_3^{al}	0.018	0.017	0.032	0.040	0.022	0.009	0.014	0.286

IV.E Substituents effect on alchemical derivatives

A significant change occurring on a property in the parent part of molecule as a result of substituent variation is called substituent effect. The substituents effect on physical and chemical properties of parent molecules can affect chemical reactivity, conformation, spectra as well as thermal properties of the substituted molecules. Substituent effects can play a major role in drug design modeling.[113,114]. For example, once having good know how about the effect of a particular substituent, then one can modify it with similar substituent to reduce the harmfulness of the chemical without significantly losing its anti-toxic nature of the drug.

Inductive and resonance (mesomeric) effects are the main electronic effects that observed due to substituent variations.[115] Inductive effect occurs as a result of an electron shift along a chain of atoms through the involvement of sigma bonds due to electronegativity or electrostatic effects. It is basically an experimentally observable effect that occurs by transmission of charge through a chain of atoms in a molecule. When the hetero atom/group attracts the electron pair towards itself, the electron withdrawing inductive (–I) effect occurs while the opposite process results in electron releasing inductive (+I) effect.[115] For instance, the substituents such as -NO_2, -CN, -F, -COOH, -Cl, -Br and -OCH_3 have –I effect where as substituents like CH_3, C_2H_5, and $(CH_3)_2CH$ possess +I effect.

The resonance (R) effect occurs from the redistribution of unsaturated (most preferably conjugated) systems through the involvement of pi-bonds. In this case, when groups adjacent to multiple bonds withdraw pi-electrons, it causes the electron withdrawing resonance (-R) effect. Groups such as –C=O, -CN, -NO_2, and -COOR are examples of this effect. In contrary, there are groups like –OCH_3, -NH_2, Cl- and –NHR which can donate electrons through resonance and thereby causes the electron releasing resonance (+R) effect.

In order to perform some analysis of influences of the conjugate bases properties on the deprotonation energy of acid, Eq.(4-1), the alchemical derivatives and the deprotonation energy has been calculated for the set of 29 H-Y molecules (Table IV-10) containing different functional groups (conjugate bases).

The relations between the group electronegativity and hardness, the electronic effects (donating or withdrawing), the increasing *s* character of the central atom of the group[116] and "alchemical" properties are discussed.

The alchemical derivative components μ^{al}, η^{al}, D^{al}, $D^{\mathrm{al,nuc}}$ and $D^{\mathrm{al,el}}$ are considered.

Table IV-10. Alchemical derivatives, the deprotonation energy and its components for H-Y molecules calculated at B3LYP/cc-pVTZ level of theory. The electronegativity and hardness at CISD/6-31++G(d,p) level [Ref.[117]].

Y	μ^{al} [a.u.]	η^{al} [a.u.]	D^{al} [a.u.]	$D^{al,nuc}$ [a.u.]	$D^{al,el}$ [a.u.]	χ [eV]	$\eta/2$ [eV]
H	-1.1074	-1.1182	0.5483	-0.7121	1.2604	-	-
CH_3	-1.1252	-1.1157	0.5673	-3.8101	4.3775	5.12	5.34
CH_2CH_3	-1.1339	-1.1340	0.5668	-5.5581	6.1250	4.42	4.96
$CHCH_2$	-1.1118	-1.1260	0.5488	-5.1142	5.6630	5.18	4.96
CCH	-1.0383	-1.0880	0.4942	-4.5553	5.0494	8.21	5.77
CHO	-1.0834	-1.1052	0.5308	-5.2470	5.7778	4.55	4.88
$COCH_3$	-1.0851	-1.1586	0.5057	-6.5630	7.0688	4.29	4.34
COOH	-0.9492	-0.9912	0.4535	-8.0594	8.5130	5.86	4.71
COCl	-1.0324	-1.1308	0.4669	-8.7523	9.2193	5.73	4.39
$COOCH_3$	-1.0749	-1.1118	0.5189	-8.4851	9.0041	5.48	-
$CONH_2$	-1.0842	-1.1074	0.5304	-5.5605	6.0909	4.67	4.42
CN	-0.9781	-1.0632	0.4465	-4.6537	5.1002	8.63	5.07
NH_2	-1.0697	-1.0290	0.5552	-4.3044	4.8596	6.16	6.04
CH_2NH_2	-1.0762	-1.0588	0.5468	-6.1437	6.6905	3.39	4.42
NO_2	-0.9488	-1.0478	0.4248	-7.9830	8.4078	7.84	4.89
OH	-1.0090	-0.9402	0.5388	-4.7512	5.2901	6.95	5.69
OCH_3	-1.0131	-0.9908	0.5177	-6.6692	7.1870	5.73	4.86
F	-0.9295	-0.8342	0.5124	-5.1653	5.6778	10.01	7.00
CH_2F	-1.1050	-1.0918	0.5591	-5.8560	6.4150	4.97	5.31
CHF_2	-1.0821	-1.0680	0.5481	-7.9662	8.5143	5.25	5.42
CF_3	-1.0512	-1.0476	0.5274	-10.057	10.5846	6.30	5.53
SiH_3	-1.1194	-1.2880	0.4753	-5.6637	6.1391	4.61	4.12
PH_2	-1.0830	-1.2688	0.4486	-6.1083	6.5570	5.05	3.96
SH	-1.0221	-1.2052	0.4195	-6.5820	7.0016	5.69	3.96
CH_2SH	-1.0417	-1.2268	0.4283	-8.1860	8.6143	4.15	4.18
Cl	-0.9535	-1.1274	0.3897	-7.0219	7.4117	7.65	4.59
CH_2Cl	-1.0911	-1.1250	0.5285	-7.3109	7.8395	4.89	4.71
$CHCl_2$	-1.0642	-1.1380	0.4952	-10.8728	11.3681	5.12	4.38
CCl_3	-1.0436	-1.1520	0.4676	-14.4570	14.9247	5.53	4.10

Electronegativity and hardness: Electronegativity is a measure of the tendency of an atom to attract a bonding pair of electrons (Pauling definition[118]) or it measures the escaping tendency of the electrons from the system that behaves like chemical potential of macroscopic thermodynamics (Mulliken definition[119]). The hardness can be viewed as a resistance of the system towards charge transfer. The electronegativity and hardness of the conjugate base depend primarily on the nature of the central atom, there is a "secondary" effect: the electronegativity and hardness of the atoms bonded to the central atom have clear effect on the total group electronegativity and hardness. Since, the proton annihilation produces free electron pair, it should be easier to remove the proton attached to more electrophilic group. In another words: the weaker the acid (lower deprotonation energy), the stronger the conjugate. This means that the decreasing tendency of the deprotonation energy with increasing electronegativity and hardness is expected. In Table IV-11, the "alchemical" properties and the group electronegativity and hardness are listed in increasing order for the sequence of the iso-valence-electronic groups.

Table IV-11. Effect of the central atom nature on the "alchemical' properties (blue colored are those molecules whose central atom is of a 2nd row while red refers for those 3rd row atoms). The group electronegativity and hardness for the radical form of group added for comparison.

property	increasing order
μ^{al}	$\mathrm{CH_3}$<$\mathrm{SiH_3}$<$\mathrm{PH_2}$<$\mathrm{NH_2}$<SH<OH<Cl<F
η^{al}	$\mathrm{SiH_3}$<$\mathrm{PH_2}$<SH<Cl<$\mathrm{CH_3}$<$\mathrm{NH_2}$<OH<F
D^{al}	**Cl<SH<$\mathrm{PH_2}$<$\mathrm{SiH_3}$<F<OH<$\mathrm{NH_2}$<$\mathrm{CH_3}$**
$D^{\mathrm{al,nuc}}$	Cl<SH<$\mathrm{PH_2}$<$\mathrm{SiH_3}$<F<OH<$\mathrm{NH_2}$<$\mathrm{CH_3}$
$D^{\mathrm{al,el}}$	$\mathrm{CH_3}$<$\mathrm{NH_2}$<OH<F<$\mathrm{SiH_3}$<$\mathrm{PH_2}$<SH<Cl
χ	$\mathrm{SiH_3}$<$\mathrm{PH_2}$<$\mathrm{CH_3}$<SH<$\mathrm{NH_2}$<OH<Cl<F
η	$\mathrm{PH_2}$<SH<$\mathrm{SiH_3}$<Cl<$\mathrm{CH_3}$<OH<$\mathrm{NH_2}$<F

The trend of the alchemical deprotonation shows a fair correlations with the electronegativity of the central atoms ($\chi_{\mathrm{C}} < \chi_{\mathrm{N}} < \chi_{\mathrm{O}} < \chi_{\mathrm{F}}$, $\chi_{\mathrm{Si}} < \chi_{\mathrm{P}} < \chi_{\mathrm{S}} < \chi_{\mathrm{Cl}}$). Within a given row of the periodic table, the alchemical potential trend is in agreement with the electronegativity trend, when for a given row in opposite, e.g. $\mu^{\mathrm{al}}_{\mathrm{CH_3}} < \mu^{\mathrm{al}}_{\mathrm{SiH_3}}; \chi_{\mathrm{CH_3}} > \chi_{\mathrm{SiH_3}}$. It should be noted that the "alchemical" properties are related to the removed proton when the group electronegativity and hardness used here are related the radical form of this group, $A\bullet$, not anionic form. It seems that the alchemical hardness ordering is more consistent with chemical intuition (showing a uniform increase from left to right in periodic Table, e.g. $\eta^{\mathrm{al}}_{\mathrm{F}} < \eta^{\mathrm{al}}_{\mathrm{OH}} < \eta^{\mathrm{al}}_{\mathrm{NH_2}} < \eta^{\mathrm{al}}_{\mathrm{CH_3}}$) while the calculated chemical hardness are not reproducing this trend. But both hardness, alchemical and chemical, show decreasing trend when going from a second to a third row central atom in the group.

Number of high electronegative atoms in group: Substituting a hydrogen atom in the methyl group by more electronegative halogen atom (F or Cl), results in the following series as are presented in Table IV-12.

Table IV-12. Effect of substituting the hydrogen atom in a methyl group by halogen atom (F or Cl).

property	increasing order
μ^{al}	$\mathrm{CH_3 < CH_2F < CH_2Cl < CHF_2 < CHCl_2 < CF_3 < CCl_3}$
η^{al}	$\mathrm{CCl_3 < CHCl_2 < CH_2Cl < CH_3 < CH_2F < CHF_2 < CF_3}$
D^{al}	$\mathbf{CCl_3 < CHCl_2 < CF_3 < CH_2Cl < CHF_2 < CH_2F < CH_3}$
$D^{\mathrm{al,nuc}}$	$\mathrm{CCl_3 < CHCl_2 < CF_3 < CHF_2 < CH_2Cl < CH_2F < CH_3}$
$D^{\mathrm{al,el}}$	$\mathrm{CH_3 < CH_2F < CH_2Cl < CHF_2 < CF_3 < CHCl_2 < CCl_3}$
χ	$\mathrm{CH_2Cl < CH_2F < CHCl_2 < CH_3 < CHF_2 < CCl_3 < CF_3}$
η	$\mathrm{CCl_3 < CH_2Cl < CHCl_2 < CH_2F < CH_3 < CHF_2 < CF_3}$

Once again, the alchemical hardness trend is “more chemically correct” than the observed chemical hardness. The alchemical results for hardness are in accordance with the fact that the experimental hardness of fluorine is higher than that of hydrogen, so the substitution of H by F in the group results in an increase in the alchemical hardness and in a decrease in D^{al}

When hydrogen is substituted by a softer atom, e,g, chlorine, the alchemical hardness and group hardness decreases. But the electronegativity of chlorine is higher than hydrogen, then the alchemical potential trend is dominant and the D^{al} increase with increasing hydrogen number. It should be noted that in the case of the alchemical derivative, the problem of overestimated value for the methyl group is not observed. In the case of the group electronegativity and hardness (the derivatives with respect to electron number), the methyl group values are higer than the value for the $\mathrm{CH_2F}$.[117]

Electronic effects: The electron donating (activating) groups such as NH_3, CH_4, and CH_3NH_2 have higher alchemical deprotonation energy than their corresponding electron withdrawing (deactivating group) like H-CN, H-CHO and H-$COOCH_3$. The larger electron withdrawing group, the greater is the inductive effect. In carbonyl compounds, oxygen atom is more electronegative than carbon atom as a result the electron density draws away from carbon and increases the bonds polarity and thereby affects the physical properties of the entire molecule. The electron releasing groups that donate an electron to the electron deficient carbonyl carbon stabilizes the molecule which in turn results in high alchemical potential. Hence, the alchemical potential for the H-CO-Y type have a decreasing order as is shown in Table IV-13.

Table IV-13. Effects of carbonyl containing substituents (H-Y) on alchemical derivatives.

property	increasing order
μ^{al}	$COCH_3$<$CONH_2$<CHO<$COOCH_3$<COCl<COOH
η^{al}	$COCH_3$<COCl<$COOCH_3$<$CONH_2$<CHO<COOH
D^{al}	**COOH<COCl<$COCH_3$<$COOCH_3$<$CONH_2$<CHO**
$D^{\mathrm{al,nuc}}$	COCl<$COOCH_3$<COOH<$COCH_3$<$CONH_2$<CHO
$D^{\mathrm{al,el}}$	CHO<$CONH_2$<$COCH_3$<COOH<$COOCH_3$<COCl
χ	$COCH_3$<CHO<$CONH_2$<$COOCH_3$<COCl<COOH
η	$COCH_3$<COCl<$CONH_2$<COOH<CHO

The effect of increasing *s* character of the central atom of the group is shown in the three following series: $D^{\mathrm{al}}_{\mathrm{CH_2CH_3}} > D^{\mathrm{al}}_{\mathrm{CCH_2}} > D^{\mathrm{al}}_{\mathrm{CCH}}$, $D^{\mathrm{al}}_{\mathrm{CH_2NH_2}} > D^{\mathrm{al}}_{\mathrm{CN}}$ and $D^{\mathrm{al}}_{\mathrm{CH_2OH}} > D^{\mathrm{al}}_{\mathrm{CHO}}$. The higher bond order the higher group electronegativity, the lower will be deprotonation energy. This can be observed in molecules such as ethane, ethylene and acetylene. The annihilation energy values calculated for these molecules are in accord with the concept of hybridization.

The overall trends of the alchemical derivative and the deprotonation energy and its components are presented in Fig. IV-8.a and Fig. IV-8.b. At panel a, the substituents are ordered in their alchemical potential increasing order (green points). It is clear from this figure that the crucial for the qualitative analysis of the substituents effect is the knowledge of the second derivative, the trend of the alchemical hardness is identical with that of the deprotonation energy. On contrary, such behavior is not observed in the case of the electron number derivatives and the group ionization energy, $IE = \chi + \eta$, (Fig. IV-8.c).

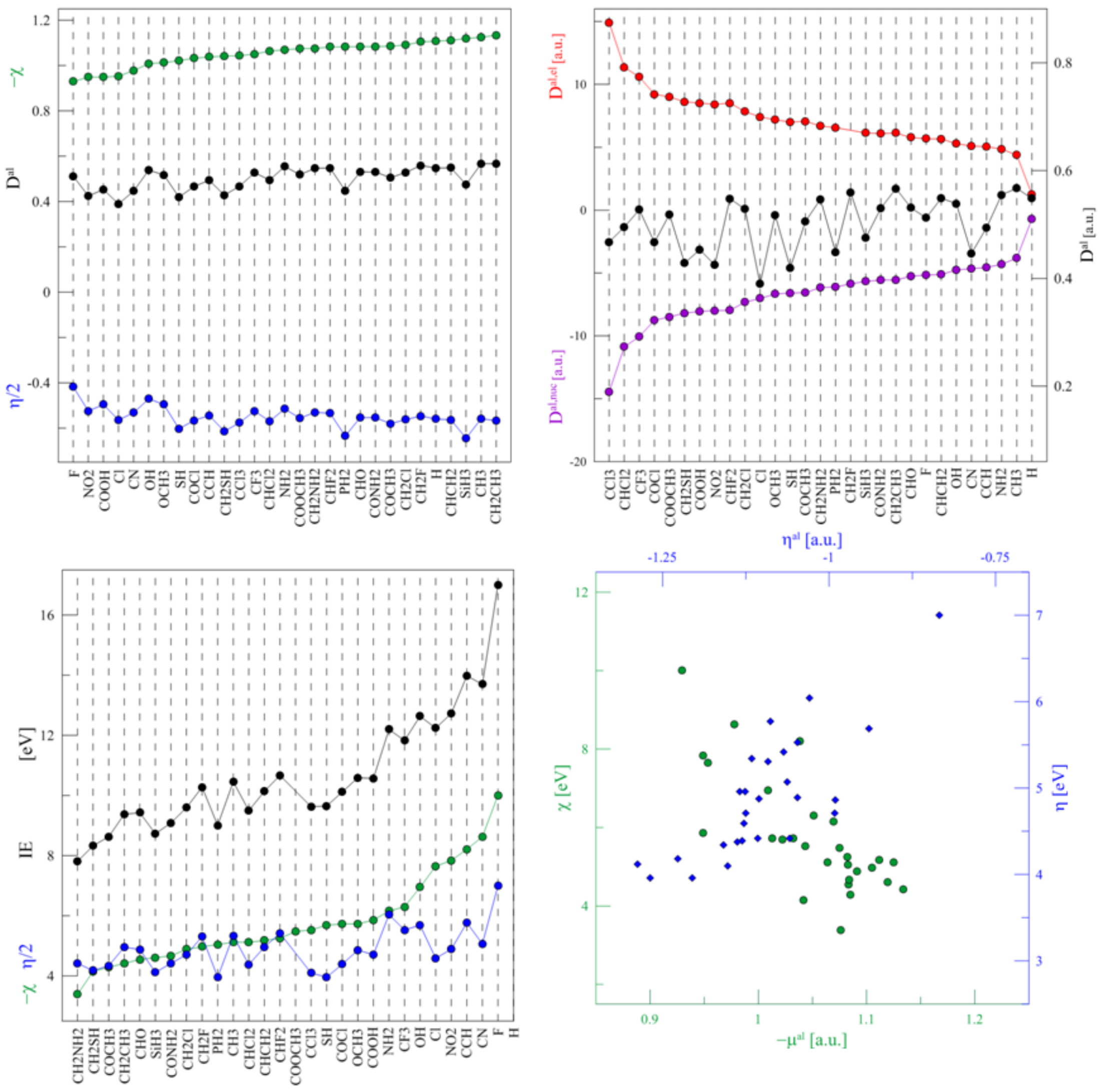

Fig. IV-8. Effect of substituents. a) alchemical deprotonation energy (black) and its derivative components (green point-first order contribution, blue points- second order contributions). Substituents are ordered in increasing order of their alchemical potential values; b) alchemical deprotonation energy (black) and its components (red electronic contribution, purple points- nuclear contributions). Substituents are ordered in the increasing order of their nuclear contribution to D^{al} ; c) Substituent Ionization energies (black) and its derivative components (green point-first order contribution, the electronegativity , blue points- second order contributions). Substituents are ordered in the increasing order of their group electronegativity; d) the relations between the derivatives with respect to electron number and the alchemical derivatives

Another intriguing result is presented in Fig. IV-8.b, the nuclear contribution to the deprotonation energy has the same but opposite order with that of electronic contribution for all the substituents, Eq.(3-32). Unfortunately, it does not mean that there is no need to evaluate $D^{\mathrm{al,el}}$ because $D^{\mathrm{al,nuc}}$ is easy to compute. The deprotonation energy does not show the monotonicity, unlike its components. Further research on this subject is planned. In Fig. IV-8.d, the relations between the derivatives with respect to electron number and the alchemical derivatives are presented (the electronegativity is the negative value of the chemical potential, the first derivative). This figure is included only for comparison (note different units).

IV.F Conclusions

In this Chapter, the usefulness of the alchemical derivatives in the prediction of the chemical properties was tested. Additionally, the basis set influence on the qualitative and quantitative accuracies of the alchemical predictions was investigated. As ”test” transmutations, the deprotonation, the transmutation of the nitrogen molecule and the substitution of the isoelectronic (B,N) units for (C,C) units and N units for C-H units were used. In all cases, the quantitative accuracy was more than satisfactory. The worst performance is observed for the deprotonation. This is connected with the slow convergence of the Taylor expansion in the case of hydrogen (compare the values of the derivatives for hydrogen and carbon in Table IV-4) and the fact that the product of the deprotonation is an anion (some vertical calculations would not lead to the carbanion energy, but rather to the energy of the complex of the radical plus an electron). The alchemical deprotonation energy (from the second order Taylor expansion) correlates well with the vertical deprotonation energy and can be used as a preliminary indicator for the experimental deprotonation energy. The introduction of the tight functions, to recover the core-core and the core-valence correlation, is crucial to achieve high quality results.

The transmutations of the benzene molecule were extensively studied. The obtained results are very good from the qualitative point of view and the quantitative point of view. In the azines case, the basis set dependence was striking. This can be explained by the occurrence of two different transmutations for which the basis set dependencies are incompatible. Therefore, this type of transmutation requires special attention and careful choice of the basis set. The results for the BN derivatives of benzene and pyrene show that the alchemical derivatives and the alchemical transmutation energies are very efficient and effective tools for the stability prediction. The splitting of the alchemical transmutation energy into electronic and nuclear parts (or into the first and second derivatives contribution) are very useful in the analysis of the dominant effect in the $(\mathrm{C,C}) \to (\mathrm{B,N})$ transmutation. It must be noted that the ratio between the derivatives and SCF calculation times for a single azaborapyrene molecule is 8:1 (the azaborapyrene number is 63). As a result, the presented method can be effectively applied to the problems such as the BN substitution in fullerenes, graphenes and similar systems.

The effect of various degrees of substituents on alchemical derivatives for H-Y type molecules depends on the (number of) electronegativity and hardness of the atom (group). In general, the weaker the acid (lower deprotonation energy), the stronger the conjugate base.

It is shown that the crucial for the qualitative analysis of the substituents effect is the knowledge of the second derivative, the trend of the alchemical hardness is identical with that of the deprotonation energy.

In summary this alchemical method has great potential for efficient and accurate scanning of Chemical Space.

Chapter V. Information and Complexity Measures in the Molecular Reactivity Studies

The use of the information theory (IT) to describe the chemical bonding and the chemical reaction was stimulated by the fact that density function is ultimately a probability distribution. The derivation of the Schrödinger equation by minimizing the Fisher information was the first application of IT concept in quantum chemistry.[120,121]. Other significant result is recovering the periodicity of atomic properties within the information theoretical framework (see Ref. [122] for review). Since then, there has been a tremendous interest to apply IT to the electronic structure theory. The concepts of uncertainty, randomness, disorder or delocalization, are the basic quantities appearing in various chemical applications. In quantum chemistry, the IT objects were used to optimize and improve basis set, to measure the amount of correlation included in a wavefunction,[123-127] to measure the similarity,[128-131] and to perform very promising IT investigation of the molecular bond[121,132-136] and reaction path.[137-140] One of the challenges is the recognition of the chemical reactivity by employing IT based measures and statistical complexity.[141,142]

As it follows from the texts concerning the complexity, there is no unique and universal definition of the complexity for arbitrary distribution.[141] Complexity is used in very different fields, although there is no general agreement about its definition. Successes in its applications suggest that the characterization of complexity cannot be univocal and must be adequate for the type of structure or process we study. In general, the initial form of complexity is designed in such a way that it vanishes for the two extreme probability distributions: corresponding to the perfect order (represented by a Dirac-delta in the shape representation) and to the maximum disorder (associated with a highly flat distribution). Recent proposals have formulated this quantity as a product of two factors, taking into account *order/disequilibrium* and *disorder/uncertainty*, respectively. One simple measure of the complexity can be defined as the product of the exponential Shannon entropy and the Fisher information (FI), the Fisher-Shannon complexity (FS), Eq.(5-9). Another complexity by replacing the FI by the Onicescu information[143,144] (OI) the López-Ruiz–Mancini–Cablet (LMC) complexity, Eq.(5-7).

The goal of the present study is two-fold: (i) answer the question, if it is preferable to use the electron density rather than the shape function as the functional argument and (ii) the recognition of the relations between the IT complexity measures, their components and the molecular reactivity. The first subject is related to the fact that all complexity measures along with their IT components are originally defined in terms of the statistical distribution functions which are normalized to unity and defined in the position space (or momentum space). This means that all properties and uncertainty relationships which are known from the statistic or the IT are valid for the $\sigma_{\mathrm{N}}(\mathbf{r})$ functional (the spinless shape), but not always for other distributions used in quantum chemistry, e.g., the spinless density. The second subject will be realized by analyzing relations between the IT measures and the DFT based chemical reactivity indices (the chemical potential and the hardness), energy components, the Pauli energy and the atomization energy.

The organization of this chapter is as follows: in Sec.V.A, the theoretical and methodological framework, the complexity measures along with their IT components and the chemical reactivity descriptors used in this work are defined and shortly discussed. In Sec. V.A.1, the molecular set chosen for study and computational details are presented. In Sec.V.B, the transferability and additivity concepts are recalled and investigated from IT point of view. The relations between the complexities components are studied in Sec. V.C, then the usability of the complexity concepts in the chemical application is deeply examined. In Sec.V.D., the correlations between the IT measures and the chemical reactivity indices are done. Finally, the conclusion related to the IT measures in the chemical reactivity study are presented in Sec.V.E. In addition in Appendix B, the relations among the IT measures in different representation (the functional dependence) are derived.

V.A Theoretical and methodological framework

A *N*-particle system is described in quantum mechanics by means of its wavefunction $\Psi(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N)$, depending on the spin-position coordinates $\mathbf{x}_i=(\mathbf{r}_i;\kappa_i)$, where $\mathbf{r}=(x,y,z)$ is a position vector and $\kappa\in\{\alpha,\beta\}$ is a spin variable. The physical and chemical properties of the systems are described by spinor electron density (Eq.(2-43)) and this density can be reduced to the spinless one (Eq.(2-87)). All these densities (including the spin components of the spinor density) will be collectively denoted by $\zeta(\mathbf{q})$ with

$$\zeta(\mathbf{q})\in\{\rho(\mathbf{x}),\rho_{\mathrm{N}}(\mathbf{r}),\rho_{\alpha}(\mathbf{r}),\rho_{\beta}(\mathbf{r})\}. \tag{5-1}$$

Their integral over the whole space (the spin-position space or the position space) is equal to the related electron number

$$\int \zeta(\mathbf{q})d\mathbf{q} = N_\zeta, \qquad N_\zeta \in \{N, N, N_\alpha, N_\beta\},\ N_\alpha + N_\beta = N. \tag{5-2}$$

Their normalized to 1 forms, the so-called the shapes, are defined as follows

$$\sigma_\zeta(\mathbf{q}) \equiv \frac{\zeta(\mathbf{q})}{N_\zeta}, \qquad \text{satisfying} \qquad \int \sigma_\zeta(\mathbf{q})d\mathbf{q} = 1. \tag{5-3}$$

The density $\zeta(\mathbf{q})$ can be characterized in different ways: as power moments $\langle r^k \rangle_\zeta$, the logarithmic moments $\langle r^k \ln r \rangle_\zeta$ or entropic moments $\langle \zeta^k \rangle_1 \equiv \langle \zeta^{k-1} \rangle_\zeta$. The symbol $\langle f(\mathbf{q}) \rangle_\zeta$ denotes the expectation value $\langle f(\mathbf{q}) \rangle_\zeta \equiv \int f(\mathbf{q})\zeta(\mathbf{q})d\mathbf{q}$. For specific values of k, these moments are physically meaningful and, at times, experimentally accessible, e.g. the electron-nucleus attraction energy which is related to the nuclear magnetic screening constant or diamagnetic screening factor, is given by $-Z_A\langle r_A^{-1} \rangle_\zeta = -Z_A \int \zeta(\mathbf{q})/|\mathbf{r} - \mathbf{R}_A| d\mathbf{q}$, $\mathbf{R}_A$ – the nucleus position.[143] Another example is the limit of the entropic moments, the Shannon entropy (SE)

$$S[\zeta] \equiv -\int \zeta(\mathbf{q}) \ln \zeta(\mathbf{q}) d\mathbf{q} = -\langle \ln \zeta \rangle_\zeta = -\lim_{k \to 1} \frac{\partial \langle \zeta^{k-1} \rangle_\zeta}{\partial k}. \tag{5-4}$$

In quantum physics, $S[\rho_\mathrm{N}]$ was introduced by von Neumann as an adaptation of the thermodynamic Gibbs entropy. For normalized argument case, it is the differential entropy $S[\sigma_\mathrm{N}]$. It is worth noting that, unlike the discrete entropy, $-\sum_i p_i \ln p_i \geq 0$, the continuous one $S[\zeta]$ can have any value in $[-\infty, +\infty]$. Any sharp peak in $\zeta(\mathbf{q})$ will tend to make negative contribution to $S[\zeta]$, whereas positive contribution are provoked by a slowly decaying tail; hence $S[\zeta]$ is a localization measure of the density. However, the differential entropy is called very often the Shannon continuous entropy (or simply SE), but it is clear that the differential entropy does not share all properties of the discrete entropy. The SE was first introduced as a way to measure the information content (or the uncertainty) in a distribution. It is a measure of localization (delocalization) or structure that exists in its functional argument. $\zeta(\mathbf{q})$[145] As a measure of the delocalization, the SE is relevant for the bonding theory.[146]

The Fisher information (FI) of the density $\zeta(\mathbf{q})$ is defined

$$I[\zeta] \equiv \left\langle \left|\nabla \ln \zeta\right|^2 \right\rangle_\zeta = \int \frac{\nabla\zeta(\mathbf{q})\cdot\nabla\zeta(\mathbf{q})}{\zeta(\mathbf{q})} d\mathbf{q}, \tag{5-5}$$

where ∇ denotes the 3-dimensional gradient. The FI, contrary to the SE, is a measure sensitive to a local spreading of the density $\zeta(\mathbf{q})$, because it is a gradient functional of $\zeta(\mathbf{q})$ The higher is this gradient, the more localized is the density (the smaller is its uncertainty of localization) and the higher is the sharpness in estimating the localization of the electrons. The sharpness, concentration or delocalization of the electronic cloud is measured by the SE as well by the FI. They give complementary descriptions of the electron localization: SE is a measure sensitive to global aspects, while the FI – to local ones. In the atomic case, the FI is a measure of density compactness, in the case of molecule, it is a measure of "peakiness". This is related to the fact that the von Weizsäcker kinetic energy (equals to $I[\rho_\mathrm{N}]/8$) is dominated by contributions of *K* electrons in atoms and molecules.[147] The FI was proposed also as the measure of the steric contribution to the total energy.[148]

The last measure, which is used here, is the Onicescu information (OI)[143,144] (the disequilibrium), the second-order entropic moment

$$D[\zeta] \equiv \left\langle \zeta^2 \right\rangle_1 = \int \zeta^2(\mathbf{q})\, d\mathbf{q} = \left\langle \zeta \right\rangle_\zeta . \tag{5-6}$$

It quantifies the departure of $\zeta(\mathbf{q})$ from equiprobability. The concept of this measure is a finer measure of dispersion distribution than that of the Shannon entropy. So far, only the mathematical aspects of this concept have been developed,[149,150] while the physical or chemical aspects were not investigated as yet.[122,151,152]

These three IT based measures are used to define two complexity measures, the LMC and FS complexities. In both, the Shannon entropy (more precisely, its exponent) is used as a measure of the total spreading of $\zeta(\mathbf{q})$. The other factor, measuring order, is the disequilibrium for the LMC complexity or the Fisher information for the FS complexities, respectively. The LMC complexity is defined[153] as

$$C_\mathrm{LMC}[\zeta] \equiv D[\zeta]H[\zeta] \tag{5-7}$$

where

$$H[\zeta] = \exp\left(S[\zeta]\right) \tag{5-8}$$

is the exponential Shannon entropy. The FS complexity is defined[154,155] as

$$C_\mathrm{FS}[\zeta] \equiv I[\zeta]J[\zeta] \tag{5-9}$$

where

$$J[\zeta]=(2\pi e)^{-1}\exp(2S[\zeta]/3)=(2\pi e)^{-1}(H[\zeta])^{2/3} \qquad (5\text{-}10)$$

is the modified exponential Shannon entropy (in three dimensional space).[156] All these definitions are valid also for the $\sigma_\zeta(\mathbf{q})$ functional argument.

The two complexity measures along with their IT components employed throughout this work were originally defined in terms of the statistical distribution functions which are normalized to unity and defined in the position space i.e. $\sigma_{\mathrm{N}}(\mathbf{r})$ (or in the momentum space). However, these measures are calculated in our paper in the spin-position space ($\mathbf{q}=\mathbf{x}$) and in the position space ($\mathbf{q}=\mathbf{r}$) in two representations: the density one and the shape one (i.e. as functionals of the density or the shape). So, the considered arguments are $\rho(\mathbf{x})$ and $\sigma(\mathbf{x})$ in the spin-position space, while $\rho_{\mathrm{N}}(\mathbf{r})$ and $\sigma_{\mathrm{N}}(\mathbf{r})$ in the position space. The relations between the IT measures and the statistical complexities for the general case are presented in Appendix B. When sets of the molecules chosen for study contain only singlet molecular systems i.e., where $\rho_\alpha(\mathbf{r})=\rho_\beta(\mathbf{r})=\rho_{\mathrm{N}}(\mathbf{r})/2$, these relations become simple. A compilation of the definitions, relationships among the various spaces and representations considered in this work is given in Table V-1. It is important to mention that the LMC and FS complexities are known to comply with the following lower bounds: $C_{\mathrm{LMC}}[\sigma_{\mathrm{N}}]\geq 1$, $C_{\mathrm{FS}}[\sigma_{\mathrm{N}}]\geq 3$ (in the spinless-shape representation) but in other representations these inequalities have to be reformulated using relations from Table V-1.

To interpret and understand the chemical nature of the IT measures, several reactivity indices are computed. The reactivity properties set used by Esquivel, and co-workers[141,142] is extended by the atomization energy, energy components and the Pauli energy. In DFT, the total electronic energy is a functional of the electron density $\rho(\mathbf{x})$ and the external potential $v(\mathbf{r})$, Eqs.(2-49) and (2-56)

$$E[\rho,v]=T_s[\rho]+E_{\mathrm{es}}[\rho]+E_{\mathrm{xc}}[\rho]+\int v(\mathbf{r})\rho(\mathbf{x})d\mathbf{x}, \qquad (5\text{-}11)$$

the sum of the kinetic energy, $T_s[\rho]$, of the non-interacting reference system, the electrostatic energy, $E_{\mathrm{es}}[\rho]$, and the exchange-correlation energy, $E_{\mathrm{xc}}[\rho]$. The ground-state energy, $E[N,v]$, is the minimum in the space of *N*-electron densities, $E[N,v]=\underset{\rho\to N}{\mathrm{Min}}\,E[\rho,v]=E[\rho^{\mathrm{gs}},v]$. The minimizer, ρ^{gs}, represents the ground-state density of the system. The sum of the electronic energy E for $v=v\left[\{Z\},\{\mathbf{R}\}\right]$ and the nuclear-nuclear repulsion energy V_{nn} is the total energy W of a system, Eq.(3-6)

$$W\left[N,\{Z\},\{\mathbf{R}\}\right]\equiv E\left[N,v\left[\{Z\},\{\mathbf{R}\}\right]\right]+V_{\text{nn}}\left[\{Z\},\{\mathbf{R}\}\right], \tag{5-12}$$

where $\{\mathbf{R}\}$ and $\{Z\}$ denote the nuclear locations and the corresponding nuclear charges. The total energy of a molecular system is often interpreted as stemming from three independent effects: steric, electrostatic, and quantum,[148] $W=W_{\text{s}}+W_{\text{es}}+W_{\text{q}}$. The electrostatic contribution is $W_{\text{es}}=E_{\text{es}}[\rho]+\int v(\mathbf{r})\rho(\mathbf{x})d\mathbf{x}$. The quantum contribution is due to the exchange-correlation and the Pauli energy, $W_{\text{q}}\left[N,\{Z\},\{\mathbf{R}\}\right]=E_{\text{xc}}[\rho]+E_{\text{P}}[\rho]$. The Pauli energy represents the portion of the kinetic energy that embodies all the effects from the antisymmetric requirement of the total wavefunction by the Pauli principle[156-158] and it is defined as[32,159,160]

$$E_{\text{P}}[\rho]=T_{\text{s}}[\rho]-T_{\text{W}}[\rho], \tag{5-13}$$

where $T_{\text{W}}[\rho]$ is the von Weizsäcker kinetic energy functional

$$T_{\text{W}}[\rho]\equiv\frac{1}{8}\int\frac{\nabla\rho(\mathbf{x})\cdot\nabla\rho(\mathbf{x})}{\rho(\mathbf{x})}d\mathbf{x}. \tag{5-14}$$

After some algebraic rearrangements, it is found that the contribution from the steric effect to the total energy equals to the Weizsäcker kinetic energy,[148] $W_{\text{s}}=T_{\text{W}}[\rho]$. It is easy to notice that the Weizsäcker kinetic energy is the rescaled Fisher information, Eq.(5-5), $T_{\text{W}}[\rho]=I[\rho]/8$.

The DFT-based chemical reactivity indices, the chemical potential and the chemical hardness are defined[32,40] as

$$\mu=-(IE+EA)/2,\qquad \eta=(IE-EA)/2, \tag{5-15}$$

where IE and EA are the vertical ionization energy and the electron affinity, respectively. In general, the chemical potential and the chemical hardness are good global descriptors of chemical reactivity. The former measures the escaping tendency of the electrons from the system and is constant through a system in equilibrium, while the latter measures the global stability of the molecule (large values of η characterize less reactive molecules–more resistive to change of electron number).

Table V-1. A compilation of the relations between $F[\zeta]$ and $F[\zeta']$ for $F \in \{S, I, D, H, J, C_{\mathrm{LMC}}, C_{\mathrm{FS}}\}$ and $\zeta, \zeta' \in \{\sigma_{\mathrm{N}}, \sigma, \rho, \rho_{\mathrm{N}}\}$, limited to spin compensated systems (where $\rho_\alpha(\mathbf{r}) = \rho_\beta(\mathbf{r}) = \rho_{\mathrm{N}}(\mathbf{r})/2$, $\sigma_\alpha(\mathbf{r}) = \sigma_\beta(\mathbf{r}) = \sigma_{\mathrm{N}}(\mathbf{r})$).

$\sigma_{\mathrm{N}} \leftrightarrow \sigma$	$\sigma \leftrightarrow \rho$	$\rho \leftrightarrow \rho_{\mathrm{N}}$	$\rho_{\mathrm{N}} \leftrightarrow \sigma_{\mathrm{N}}$
$S[\sigma] = S[\sigma_{\mathrm{N}}] + \ln 2$	$S[\rho] = N(S[\sigma] - \ln N)$	$S[\rho_{\mathrm{N}}] = S[\rho] - N \ln 2,$	$S[\sigma_{\mathrm{N}}] = S[\rho_{\mathrm{N}}]/N + \ln N$
$I[\sigma] = I[\sigma_{\mathrm{N}}]$	$I[\rho] = N\, I[\sigma]$	$I[\rho_{\mathrm{N}}] = I[\rho]$	$I[\sigma_{\mathrm{N}}] = I[\rho_{\mathrm{N}}]/N$
$D[\sigma] = D[\sigma_{\mathrm{N}}]/2$	$D[\rho] = N^2 D[\sigma]$	$D[\rho_{\mathrm{N}}] = 2D[\rho]$	$D[\sigma_{\mathrm{N}}] = D[\rho_{\mathrm{N}}]/N^2$
$H[\sigma] = 2H[\sigma_{\mathrm{N}}]$	$H[\rho] = (H[\sigma]/N)^N$	$H[\rho_{\mathrm{N}}] = 2^{-N} H[\rho]$	$H[\sigma_{\mathrm{N}}] = N(H[\rho_{\mathrm{N}}])^{-N}$
$J[\sigma] = 2J[\sigma_{\mathrm{N}}]$	$J[\rho] = (2\pi e)^{1-N} N^{2N/3} (J[\sigma])^N$	$J[\rho_{\mathrm{N}}] = 2^{-2N/3} J[\rho]$	$J[\sigma_{\mathrm{N}}] = N^{2/3} (2\pi e)^{(N-1)/N} (J[\rho_{\mathrm{N}}])^{1/N}$
$C_{\mathrm{LMC}}[\sigma] = C_{\mathrm{LMC}}[\sigma_{\mathrm{N}}]$	$C_{\mathrm{LMC}}[\rho] = N^{2-N} (D[\sigma])^{1-N} (C_{\mathrm{LMC}}[\sigma])^N$	$C_{\mathrm{LMC}}[\rho_{\mathrm{N}}] = 2^{1-N} C_{\mathrm{LMC}}[\rho]$	$C_{\mathrm{LMC}}[\sigma_{\mathrm{N}}] = N^{-1} (D[\rho_{\mathrm{N}}])^{N+1} (C_{\mathrm{LMC}}[\rho_{\mathrm{N}}])^{-N}$
$C_{\mathrm{FS}}[\sigma] = 2^{2/3} C_{\mathrm{FS}}[\sigma_{\mathrm{N}}]$	$C_{\mathrm{FS}}[\rho] = \dfrac{N^{(2N+3)/3}}{(2\pi e)^{N-1}} (I[\sigma])^{-N} (C_{\mathrm{FS}}[\sigma])^N$	$C_{\mathrm{FS}}[\rho_{\mathrm{N}}] = 2^{-2N/3} C_{\mathrm{FS}}[\rho]$	$C_{\mathrm{FS}}[\sigma_{\mathrm{N}}] = \dfrac{(2\pi e)^{(N-1)/N}}{N^{1/3}} (I[\rho_{\mathrm{N}}])^{-1/N} (C_{\mathrm{FS}}[\rho_{\mathrm{N}}])^{1/N}$

In addition to these global reactivity indices, the atomization energy (activation energy) will be used as the molecular stability descriptor. It is simply defined as the energy change that accompanies the total separation of all atoms in a molecule

$$\Delta W_M = \sum_{A \in M} W_A - W_M \qquad (5\text{-}16)$$

where W_M – the total energy of molecule, W_A – the total energy of an atom A.

V.A.1 Molecular set and computational details Theoretical and methodological framework

The molecular set chosen for the study includes different types of molecules with various functional groups (see Scheme V-1). This set can be divided into three subsets namely the alkane-alkene-alkyne set (A3 set, 39 molecules,), X–Y molecule set (XY set, 190 molecules) and the *ortho*-, *meta*-, *para*- substituted benzene derivatives (*d*-XY set, 170 molecules at each position). These sets represent a variety of chemical organic systems, including alcohol (-OH), carboxylic acid (-COOH), aldehyde (-CHO), ketone ($-COCH_3$), ether ($-CH_2OCH_3$), ester ($-COOCH_3$) and amide ($-CONH_2$), for both non-aromatic and aromatic molecules. These sets can be also split into a subset of molecules with non-hydrogen atom from the II row only, from the III row only and the remaining molecules (with atoms from the II and III rows simultaneously). These subsets will be denoted by II, III, and m. The used total set is large enough and diverse to improve the understanding of molecular complexity and to generalize obtained conclusions. All calculations were done for spin singlet systems using GAMESS program package.[161] As the exchange-correlation functional, the B3LYP functional was used, known as the most widely used functional in trouble-free situations (covalent interaction, excluding π-π stacking and no transition metal bonds).[89,90] The cc-pVTZ basis set[92] was used as the optimal balance choice between accuracy and computation time. In addition, the XY set was tested with different basis sets (aug-cc-pVTZ) and different approximate exchange-correlation functionals, to see the impact on atomic and molecular values of Shannon entropy and Fisher information. No significant dependence on the choice of basis sets and functional forms has been observed.

For each molecule different information and complexity measures, i.e. S, I, D, C_{LMC} and C_{FS} were calculated as the spinor-density functionals as well as the spinless-shape functionals all in the configuration space. The ionic forms of molecules and atomic data were calculated at the same level. The atomic information measures were calculated for the equi-ensemble density (the average over the spin and spatial degeneracy, see example in Table

V-2) as it was presented in Sec.II.D. Density matrices. All quantities are given in atomic units throughout this section.

$$\mathrm{C}_n\mathrm{H}_{2n+2}\,(\blacktriangle),\mathrm{C}_n\mathrm{H}_{2n}\,(\blacktriangledown),\ \mathrm{C}_n\mathrm{H}_{2n\text{-}2}\,(\blacktriangleright),\qquad 3\le n\le 10 \qquad \text{A3}$$

$$X-Y,\ \{X,Y\}\in\left\{\begin{matrix}-\mathrm{H},-\mathrm{CH_3},-\mathrm{NH_2},-\mathrm{OH},-\mathrm{F},-\mathrm{SiH_3},-\mathrm{PH_2},-\mathrm{SH},\\ -\mathrm{Cl},-\mathrm{CONH_2},-\mathrm{OCH_3},-\mathrm{CHO},-\mathrm{COCH_3},-\mathrm{C_6H_5},\\ -\mathrm{COOH},-\mathrm{COOCH_3},-\mathrm{CH_2OCH_3},-\mathrm{CN},-\mathrm{NO_2}\end{matrix}\right\}(\bullet) \qquad \text{X–Y}$$

$$d-X-\mathrm{C_6H_4}-Y,\qquad d\in\left\{ortho(\square),meta(\text{▯}),para(\blacksquare)\right\} \qquad \text{d-XY}$$

Scheme V-1. Molecular sets for IT components and complexity measures. The symbols used in figures are here in parenthesis. The set names are on the right. The color coding used in figures throughout this chapter uses: red – the molecules with non-hydrogen atoms from II row only and the hydrogen molecule; green – the molecules with non-hydrogen atoms from III row only; blue – the remaining molecules.

Table V-2. IT-measures (SE and FI) of carbon atom for different types of densities calculated at B3LYP/cc-pVTZ. All are in a.u.

Angular momentum		SE				FI			
Orbital	Spin	$\rho_\alpha(\mathbf{r})$	$\rho_\beta(\mathbf{r})$	$\rho(\mathbf{x})$	$\rho_\mathrm{N}(\mathbf{r})$	$\rho_\alpha(\mathbf{r})$	$\rho_\beta(\mathbf{r})$	$\rho(\mathbf{x})$	$\rho_\mathrm{N}(\mathbf{r})$
$L_z=0$	$S_z=0$	5.91	5.91	11.82	7.67	128.05	128.05	256.10	256.10
	$S_z=+1$	8.25	3.02	11.27	7.67	128.04	129.78	257.82	256.10
$\bar{L}_z=0$	$S_z=0$	5.96	5.96	11.92	7.76	127.56	127.56	255.12	255.1
	$S_z=+1$	8.40	3.02	11.42	7.76	126.53	129.77	256.30	255.12
	$\bar{S}_z=0$	5.96	5.96	11.92	7.76	127.56	127.56	255.12	255.12

V.B Additivity and transferability of the IT measures

As the argument in discussion about whether more information is revealed using the shape function or using the electron density, the transferability and additivity concepts will be used. They are central to chemistry: the idea that every property of a molecule is approximated by the sum of the contributions from each of its constituent atoms or groups. The observation of "experimental group additivity" requires that in addition to the properties of the groups being additive, the group and its properties have to be transferable from one molecule to another (different) molecule.[162,163] These concepts of transferability and additivity will be tested on the example of separate atoms, the $\mathrm{CH_2}$ group and $\mathrm{C_6H_4}$ group contributions.

The atomic additivity and transferability is investigated using the linear fitting without a constant

$$\tilde{F}\left[\omega_M\right]=a\sum_{A\in M}F\left[\omega_A^0\right],\qquad \begin{array}{l} F\in\{S,I,D\}, \\ \omega(\mathbf{q})\in\{\sigma_{\mathrm{N}}(\mathbf{r}),\sigma(\mathbf{x}),\rho_{\mathrm{N}}(\mathbf{r}),\rho(\mathbf{x})\}, \end{array}, \tag{5-17}$$

to the molecular value $F\left[\omega_M\right]$, for a set of molecules. All three measures used here in the spinor-density representation, $\rho(\mathbf{x})$, show a very high-accuracy linear relation between the molecular measure value and the sum of the atomic measure values, the linear fitting data are presented in Table V-3. Combination of this fit with the atomization measure, which is defined as

$$\Delta F[\omega]=\sum_{A\in M}F\left[\omega_A^0\right]-F\left[\omega_M\right], \tag{5-18}$$

yields

$$\Delta F[\omega]=(1-a)\sum_{A\in M}F\left[\omega_A^0\right]+\left(\tilde{F}\left[\omega_M\right]-F\left[\omega_M\right]\right), \tag{5-19}$$

where ω_A^0 is the ω-type argument for a free atom.

Table V-3. Fitting of the molecular value vs. the sum of atomic values for the Shannon entropy, the Fisher information, the Onicescu information, all in the spinor-density representation, as a linear function, $F\left[\rho_M\right]=a\left(\sum_{A\in M}F\left[\rho_A\right]\right)$, and the Fisher information approximated by *K*-shells model (see Eq.(5-20)). The mean absolute error (MAE) and the mean absolute percentage error (MAPE). $\mathrm{R}^2=1.000$ for each fitting.

$F[\rho]$	a	MAE	MAPE
$S[\rho]$	0.788	0.722	1.441
$I[\rho]$	0.974	25.638	0.926
$D[\rho]$	0.997	0.720	0.339
I_K	0.880	67.464	2.345

This means that if the slope value is close to one, the atomization value is close to a fitting error, and observed transferability is the result of compensatory transferability wherein the changes in the properties of one atom/group are compensated for by equal but opposite change in the properties of the adjoining atom/group.[163] In Fig. V-1, the relation between the molecular value and the atomization value are presented in left panels. The atomization values for all measures are positive. In the case of the FI and the OI, the slope is close to one (see Table V-3). Small atomization values in comparison with the molecular values for the FI and the OI confirm that the compensating changes are small. Using the fact that the major contribution to the Weizsäcker kinetic energy comes from the *K*-shell electrons for neutral atoms,[147] the *K*-shells model to approximate the FI is tested (see Table V-3)

$$I_K = 4\sum_A n_A^K Z_A^2, \qquad n_A^K = \{1 \text{ for } A = H;\ 2 \text{ for } A \neq H\}, \tag{5-20}$$

where n_A^K is the *K*-shell occupation number (one for hydrogen atom, two for the rest) and Z_A is the nuclear charge of A atom. The high-accuracy of this model confirms the major contribution to the FI coming from core electron density. Similar model for the OI, $D_K = \sum_A n_A^K Z_A^3 / 8\pi$, based on the hydrogenic atoms[164] produces large errors.

In the case of the SE, the slope significantly differs from one and, what is more, the $\Delta S[\rho]$ can be fitted as the linear function of $S[\rho]$ (see the black line in ΔS vs S). The slope value ($a = 0.788$) should be interpreted as the ratio of the average entropy of the atoms in molecule to the average entropy of free atoms which are constituents of molecules used in this work. It depends on the used set, e.g. for the X–Y-III set and d-XY-III, it is equal 0.806 and 0.777, respectively, but still with high correlation. This behavior is due to the fact that during bond formation, the major changes in the density are observed for the valence part.

As can be noted from the right panels of Fig. V-1, compared with the relations given in Table V-1, the transferability and additivity of atoms shown by the IT measures in the spinor-density representation disappear in the spinless-shape representation. The highly accurate linear relation between the SE of the molecule and the sum of the atomic SEs in the ρ representation (see Table V-3), is lost in the ρ_{N} representation as can be easily shown using transformation from Table V-1,

$$S[\rho_M] = a\left(\sum_{A\in M} S[\rho_A]\right) \quad \Rightarrow S\left[\rho_{\mathrm{N}}^M\right] = a\left(\sum_{A\in M} S\left[\rho_{\mathrm{N}}^A\right]\right) + (a-1) N_M \ln 2 \,. \tag{5-21}$$

The transferability of the methylene group, CH_2 has been extensively studied.[163,165-167] In Fig. V-2, the data for the A3 set is presented in both representations. The most remarkable feature that may be observed from these figures is that the density representation (left panels) preserves the group transferability and the size extensiveness. For all three measures, we observed an increasing value with increasing molecular size. Using the electron number as a molecule size index, the excellent fittings to linear relationships are obtained

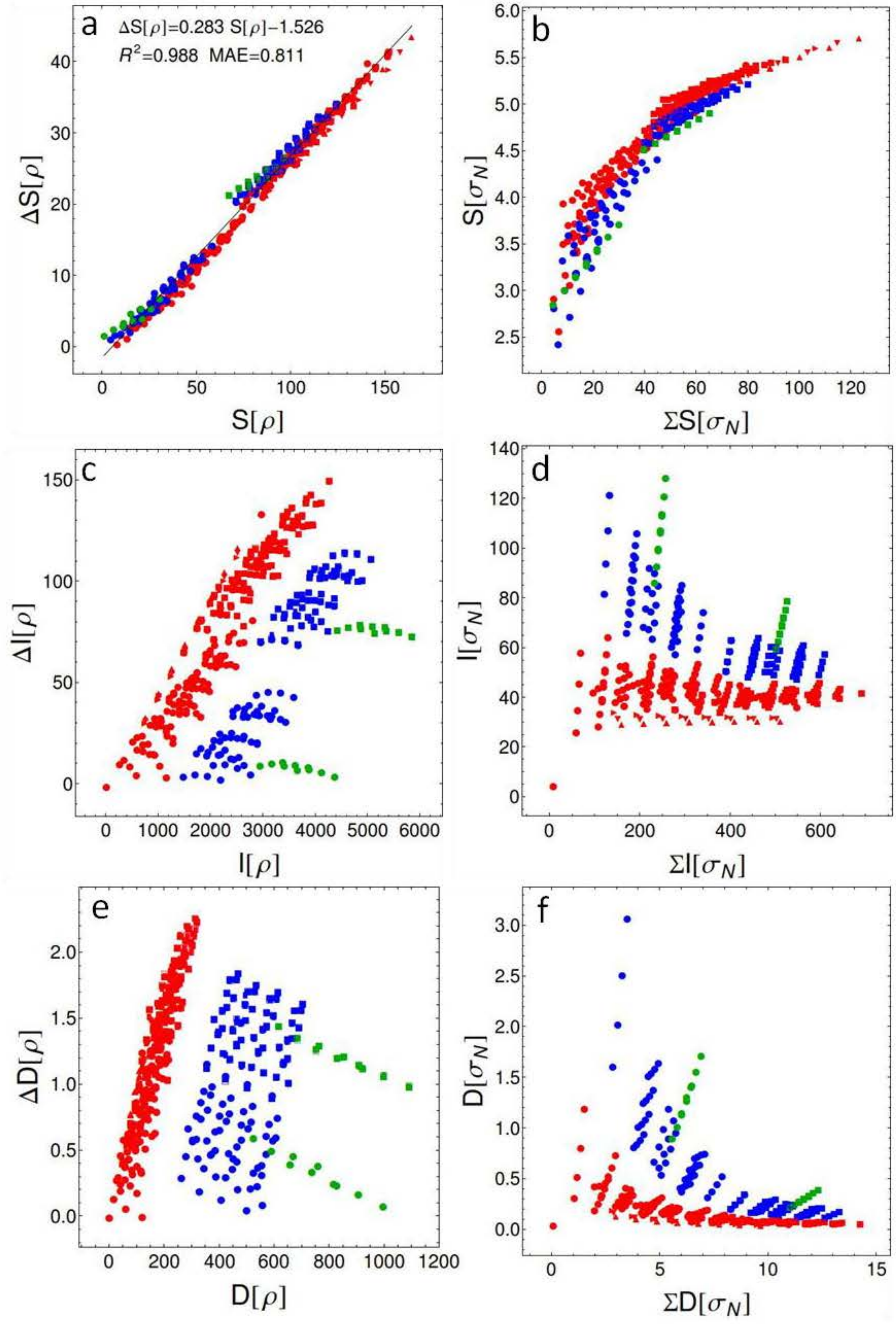

Fig. V-1. The IT measures of the atomization in the spinor density representation (left panels) and the atomic additivity of the IT measures in the shape representation (right panel). For axis labels see the text, for symbols used in figures see Scheme V-1.

(see Table V-4). Almost the same slopes of subsets imply for the given measure the conservation of the methylene properties – the same $8a$ contribution to this measure. Zooming in the molecule values, we find for the hydrocarbons with the same carbon number the following trend: from alkane to alkene to alkyne, their value of three measures used here decrease.Using transformation from the ρ representation to the σ_{N} representation (Table V-1), the $\rho \to \rho_{\mathrm{N}} \to \sigma_{\mathrm{N}}$ mapping), we obtain for the σ_{N}

$$S\left[\sigma_{\mathrm{N}}\right] = \left(a - \ln 2\right) + \frac{b}{N} + \ln N\,, \qquad I\left[\sigma_{\mathrm{N}}\right] = a + \frac{b}{N}, \qquad D\left[\sigma_{\mathrm{N}}\right] = \frac{2a}{N} + \frac{2b}{N^2}\,, \tag{5-22}$$

where a and b are the coefficients of the linear fitting $f\left[N\right] = aN + b$ for ρ representation (see Table V-4). The result for the SE is consistent with the approximate linearity of the sum of Shannon entropies in the position and the momentum space with $\ln N$ observed empirically for atoms.[122] The large N limits are associated with the asymptotic behavior of the OI at zero value and of the FI at the a values, e.g. 31.391 a.u in the alkenes case. As can be noticed form Fig. V-2 and Eq.(5-22), with increasing size N of hydrocarbons the $I\left[\sigma_{\mathrm{N}}\right]$ values for the alkanes and alkynes become finally the same as the alkenes values. Very often the SE and the FI are considered to be the different sides of the same coins. But after examination of the trends between alkanes, alkenes and alkynes in the shape representation, we conclude that the FI value for alkynes is always higher than for alkenes and for alkenes is higher than for alkanes irrespective of the number of carbon atoms they contain. For a given number of carbon atoms, the entropy trend is opposite to the disequilibrium trend. This observation is in contradiction with the meaning of both measures, because the OI is considered to be finer measure of dispersion distribution than the SE. The FI is interpreted as the steric effect measure. The FI results for the A3 set are inconsistent with the expectation that the steric effect needs to be extensive in size (the FI for alkenes are almost constant with increasing size in the shape representation). All discussed contradictions are observed for the shape representations only.

Table V-4. The Shannon entropy, the Fisher information, the Onicescu information, all in the spinor-density representation, fitted as a linear function of the electron number, $f\left[N\right] = aN + b$, and the mean absolute error, for the subsets of the A3 set. All in a.u.

	$S\left[\rho\right]$		$I\left[\rho\right]$		$D\left[\rho\right]$	
alkanes	$1.946N + 4.738$	0.146	$31.370N - 51.710$	0.257	$1.973N - 3.909$	0.018
alkenes	$1.955N + 1.578$	0.058	$31.390N + 4.091$	0.170	$1.972N + 0.045$	0.008
alkynes	$1.961N - 1.099$	0.020	$31.400N + 59.990$	0.096	$1.972N + 3.947$	0.007

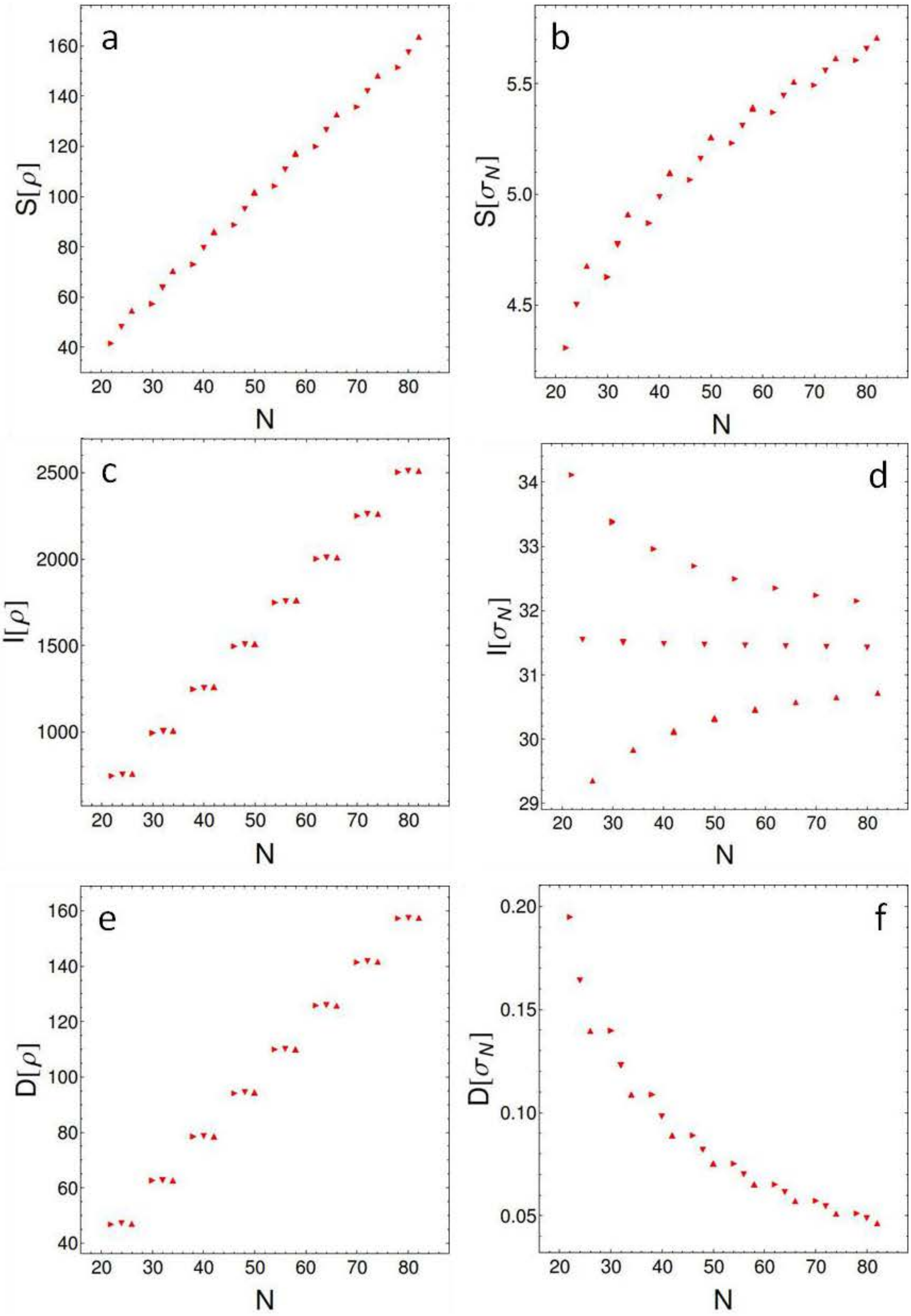

Fig. V-2. The IT measures vs. electron number for the hydrocarbons (the A3 set) in the spinor density representation (left panels) and the shape representation (right panel). See Fig. V-1 for details.

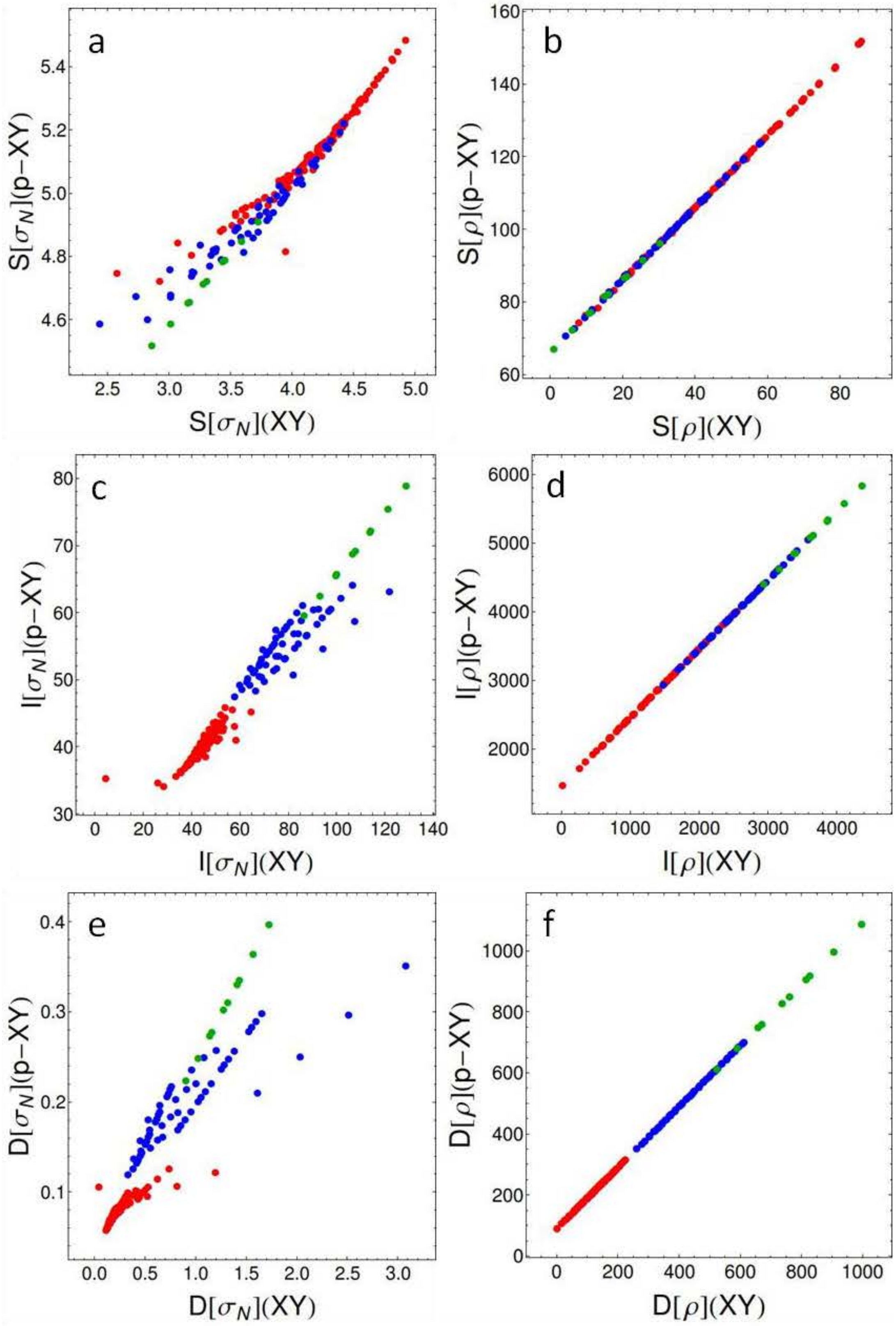

Fig. V-3. The $p-C_6H_6$ group transferability of the IT measures between the X–Y set and *p*-XY set in the spinor density representation (right panels) and the shape representation (left panel).

Another example, which illustrate that the conclusions derived from the shape representation are inconsistent with the group transferability, is presented in Fig. V-3. The group additivity and transferability is tested on the examples of the C_6H_4 group present in the d-XY set as compared with the X–Y set. This group transferability was tested using the linear fitting

$$\tilde{F}\left[\rho_{d-XY}\right]=a_{\mathrm{F}}F\left[\rho_{XY}\right]+b_{\mathrm{F}} \qquad (5\text{-}23)$$

where ρ_{XY} and ρ_{d-XY} are the densities of the X–Y molecule and its benzene derivative. These relations in the case of substituents at the *para* position, in the spinor-density representation, are perfectly linear with $R^2=1.000$ (Fig. V-3, right panels)

$$\begin{aligned}
&S\left[\rho_{p-\mathrm{XY}}\right]=0.997\ S\left[\rho_{\mathrm{XY}}\right]+66.409, \quad \text{MEA=0.160,}\\
&I\left[\rho_{p-\mathrm{XY}}\right]=1.000\ I\left[\rho_{\mathrm{XY}}\right]+1477.000, \ \text{MEA=0.937,}\\
&D\left[\rho_{p-\mathrm{XY}}\right]=1.000\ D\left[\rho_{\mathrm{XY}}\right]+94.430, \quad \text{MEA=0.034.}
\end{aligned} \qquad (5\text{-}24)$$

This demonstrates additivity of X and Y groups, in the X–Y molecules and the *p*-XY molecules, and transferability of the $p-\mathrm{C_6H_4}$ group. The intercepts in Eq.(5-24) are the IT measure values for the free p-$\mathrm{C_6H_4}$ group. Similar to Eq.(5-24), perfectly linear relations ($R^2=1.000$) are obtained in the case of substituents at the *meta* and *ortho* position. The difference between the free m-$\mathrm{C_6H_4}$ and the p-$\mathrm{C_6H_4}$ are -0.022, 0.291, 0.001 for SE, FI and OI, respectively. For the o-$\mathrm{C_6H_4}$, these values are -0.072, 0.139 and -0.006. Such small differences indicate the transferability of the $\mathrm{C_6H_4}$ group with respect to various positions for binding substituent.

Recalling the relation between the IT measures in the shape representation and the spin-density representation, $S[\rho]=N\left(S[\sigma_N]+\ln 2-\ln N\right)$, $I[\rho]=N\,I[\sigma_{\mathrm{N}}]$ and $D[\rho]=N^2\,D[\sigma_N]/2$, it is obvious why each linear relation, Eq.(5-24), is not observed in the shape representation (see Fig. V-3, left panels).

V.C The information theory planes and complexities

The concept of the complexity seems to be very attractive in the chemical reactivity investigations. To the best of our knowledge, there is only one example of the complexity analysis from the molecular reactivity point of view.[141,142] To begin investigation, it is interesting to analyze the relation between the complexities components. This is done in Fig.

V-4 by means of the information planes. On the $D-H$ plane, Fig. V-4.a, we notice an approximate $D\sim 1/H$ relation between the entropy power, $H[\sigma_N]$, and the disequilibrium, $D[\sigma_N]$ for a subset of molecules without the III row atoms (the red points, the unfit point is the hydrogen molecule), yielding values close to 16 for the $C_{\mathrm{LMC}}=DH$ complexity (see the $C_{\mathrm{FS}}-C_{\mathrm{LMC}}$ plane, Fig. V-4.d). A larger dispersion around $1/H$ behavior for the molecules containing the III-row atoms (blue and green points) results in a wide range of the $C_{\mathrm{LMC}}[\sigma_N]$ values. Although the values for the benzene derivatives, the *d*-XY set (blue and green squares) are located at $H>85$, while their X–Y set counterparts (blue and green circles) are at $H<85$, nevertheless this results in the same wide range of C_{LMC} values (Fig. V-4.d).

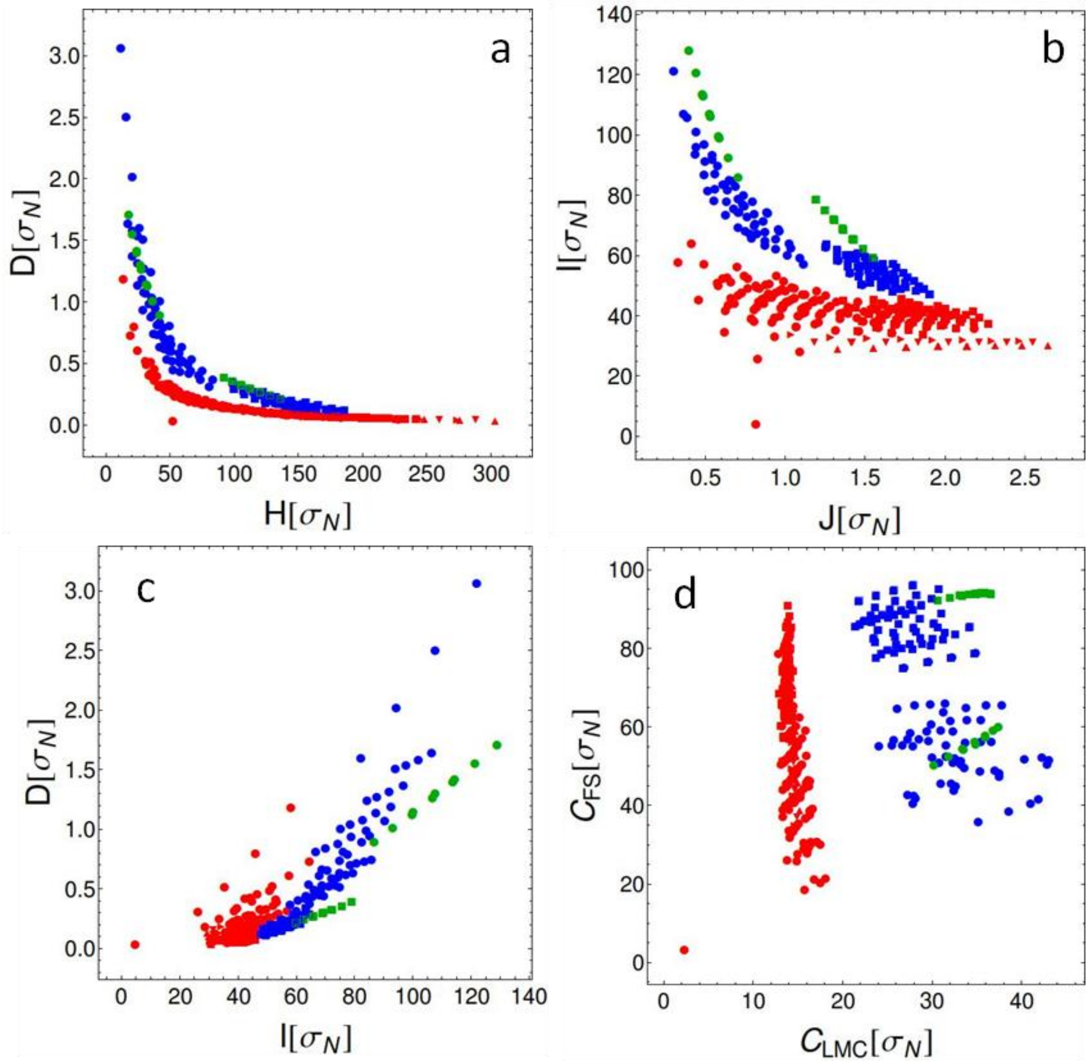

Fig. V-4. The IT measures and complexities planes in the shape representation (panels a-d).

In the case of the C_{FS} complexity, $C_{\mathrm{FS}} = IJ$, the range of the values for the molecules indicated by the red points in Fig. V-4.d is significantly wider than in the C_{LMC} case, but for the molecules marked by the green squares (the benzene derivative with heavy atom in substituent group only from III rows) is quite narrow. The inconsistency between two complexities can be easily shown on the example of the A3 set. Insertion of the relations from Eq.(5-22) into the complexities definitions yields

$$\begin{aligned} C_{\mathrm{LMC}}\left[\sigma_{\mathrm{N}}\right] &= \left(\frac{a_{\mathrm{D}}N + b_{\mathrm{D}}}{N}\right)\exp\left(a_{\mathrm{S}} + \frac{b_{\mathrm{S}}}{N}\right), \\ C_{\mathrm{FS}}\left[\sigma_{\mathrm{N}}\right] &= \frac{\left(a_{\mathrm{I}}N + b_{\mathrm{I}}\right)N^{2/3}}{2^{1/3}\pi e N}\exp\left(\frac{2}{3}\left(a_{\mathrm{S}} + \frac{b_{\mathrm{S}}}{N}\right)\right), \end{aligned} \tag{5-25}$$

where a and b are the fitting coefficients for a given measure, $F \in \{S, I, D\}$, Table V-4. As the hydrocarbon size N increases, the C_{LMC} complexity decreases while the C_{FS} complexity increases, two opposite effects (Fig. V-5.a).

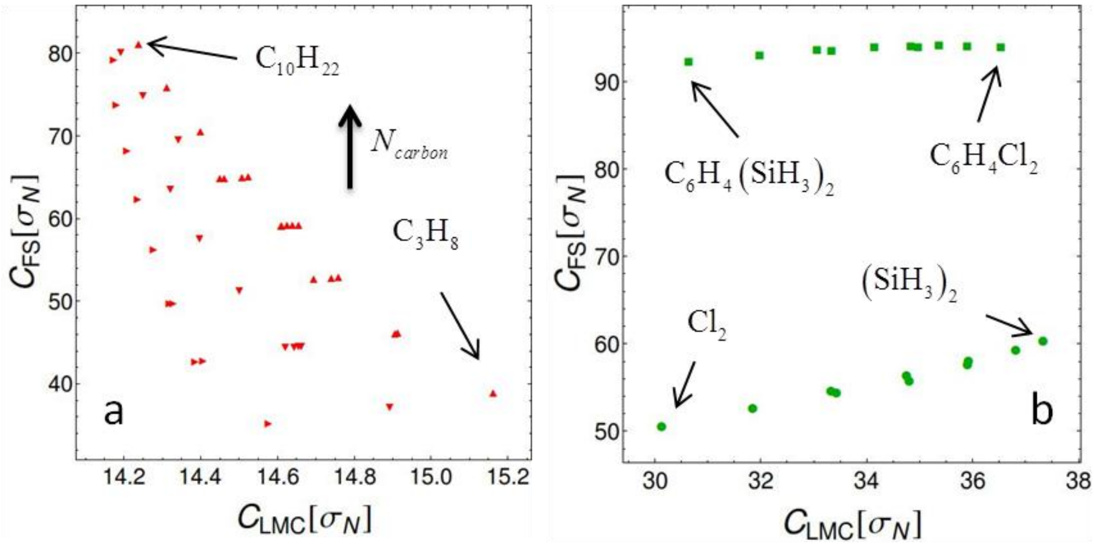

Fig. V-5. The inconsistence between $C_{\mathrm{LMC}}\left[\sigma_{\mathrm{N}}\right]$ and $C_{\mathrm{FS}}\left[\sigma_{\mathrm{N}}\right]$ for the A3 set (panel a) and for the X–Y-III and d-XY-III sets (panel b). See Fig. V-1 for details.

Another example is the complexity for the X–Y-III set and the d-X–Y-III set, in Fig. V-5.b. The molecules from the X–Y-III set are isoelectronic molecules with $N = 34$ electrons. Assuming that the A molecule has higher complexity than the B molecule, the following relation occurs for the C_{LMC} complexity

$$\begin{aligned} C_{\mathrm{LMC}}\left[\sigma_{\mathrm{N}}^{A,\mathrm{XY}}\right] > C_{\mathrm{LMC}}\left[\sigma_{\mathrm{N}}^{B,\mathrm{XY}}\right] &\Leftrightarrow \left(S\left[\sigma_{\mathrm{N}}^{A,\mathrm{XY}}\right]-S\left[\sigma_{\mathrm{N}}^{B,\mathrm{XY}}\right]\right) > \ln\frac{D\left[\sigma_{\mathrm{N}}^{B,\mathrm{XY}}\right]}{D\left[\sigma_{\mathrm{N}}^{A,\mathrm{XY}}\right]} \\ &\Leftrightarrow \left(S\left[\rho_{\mathrm{XY}}^{A}\right]-S\left[\rho_{\mathrm{XY}}^{B}\right]\right) > N_{\mathrm{XY}}\ln\frac{D\left[\rho_{\mathrm{XY}}^{B}\right]}{D\left[\rho_{\mathrm{XY}}^{A}\right]}, \end{aligned} \tag{5-26}$$

and for the C_{FS} complexity

$$\begin{aligned} C_{\mathrm{FS}}\left[\sigma_{\mathrm{N}}^{A,\mathrm{XY}}\right] > C_{\mathrm{FS}}\left[\sigma_{\mathrm{N}}^{B,\mathrm{XY}}\right] &\Leftrightarrow \frac{2}{3}\left(S\left[\sigma_{\mathrm{N}}^{A,\mathrm{XY}}\right]-S\left[\sigma_{\mathrm{N}}^{B,\mathrm{XY}}\right]\right) > \ln\frac{I\left[\sigma_{\mathrm{N}}^{B,\mathrm{XY}}\right]}{I\left[\sigma_{\mathrm{N}}^{A,\mathrm{XY}}\right]} \\ &\Leftrightarrow \left(S\left[\rho_{\mathrm{XY}}^{A}\right]-S\left[\rho_{\mathrm{XY}}^{B}\right]\right) > \frac{3}{2}N_{\mathrm{XY}}\ln\frac{I\left[\rho_{\mathrm{XY}}^{B}\right]}{I\left[\rho_{\mathrm{XY}}^{A}\right]}. \end{aligned} \tag{5-27}$$

Using the data from Table V-5, the lowest complexity in the X–Y-III set is for chlorine molecule and the highest for disilane (the complexity for the X–Y-III molecules decreases). The relations from Eqs.(5-26) and (5-27) for the benzene derivative at position *para* can be rewritten using Eq.(5-24)

$$\begin{aligned} &C_{\mathrm{LMC}}\left[\sigma_{\mathrm{N}}^{A,p-\mathrm{XY}}\right] > C_{\mathrm{LMC}}\left[\sigma_{\mathrm{N}}^{B,p-\mathrm{XY}}\right] \Leftrightarrow \\ &\qquad 0.997\left(S\left[\rho_{\mathrm{XY}}^{A}\right]-S\left[\rho_{\mathrm{XY}}^{B}\right]\right) > N_{p-\mathrm{XY}}\ln\frac{D\left[\rho_{\mathrm{XY}}^{B}\right]+94.430}{D\left[\rho_{\mathrm{XY}}^{A}\right]+94.430}, \end{aligned} \tag{5-28}$$

$$\begin{aligned} &C_{\mathrm{FS}}\left[\sigma_{\mathrm{N}}^{A,p-\mathrm{XY}}\right] > C_{\mathrm{FS}}\left[\sigma_{\mathrm{N}}^{B,p-\mathrm{XY}}\right] \Leftrightarrow \\ &\qquad 0.997\left(S\left[\rho_{\mathrm{XY}}^{A}\right]-S\left[\rho_{\mathrm{XY}}^{B}\right]\right) > \frac{3}{2}N_{p-\mathrm{XY}}\ln\frac{I\left[\rho_{\mathrm{XY}}^{B}\right]+1477.000}{I\left[\rho_{\mathrm{XY}}^{A}\right]+1477.000}, \end{aligned} \tag{5-29}$$

and with data from Table V-5, the lowest complexity is for dichlorobenzene and the highest for disilylbenzene (the complexity for the p-X–Y-III molecules has the opposite ordering as compared to the X–Y-III molecules). With almost the same range of C_{LMC} for both sets, the C_{FS} values are well separated, Fig. V-5.b.

In the spinor density representation the IT planes possess more pattern and organization information than those presented in Fig. V-4. In Fig. V-6.a-c, the information planes between the three IT based measures are plotted. We can observe similar patterns between the X–Y set and the d-XY set (circles and squares, respectively), which can be interpreted as the $\mathrm{C_6H_4}$ group transferability effect. A linear behavior can be observed for molecules with the same X and different Y group and are isoelectronic molecules, e.g. green points which are related to the X–Y-III and d-XY-III set molecules. For such molecules, ordered as is written in Table V-5, we observe the decrease in uncertainty, $S[\rho]$, accompanied by increase of order and organization, $D[\rho]$ and $I[\rho]$, respectively. We can see that for the molecular set used in this work, the hydrocarbons from the A3 set and hydrogen molecule are the lower bounds in all IT planes. Comparing the planes presented in Fig. V-6, it can be observed that the $S-I$ plane provides "richer" information about the pattern, organization, similarity of molecules than the other planes. Note that the strong correlation between SE and the FI information observed for the atoms[168] and the hydrocarbons,[169] does not reflect a general relation (see Fig. V-6. a).

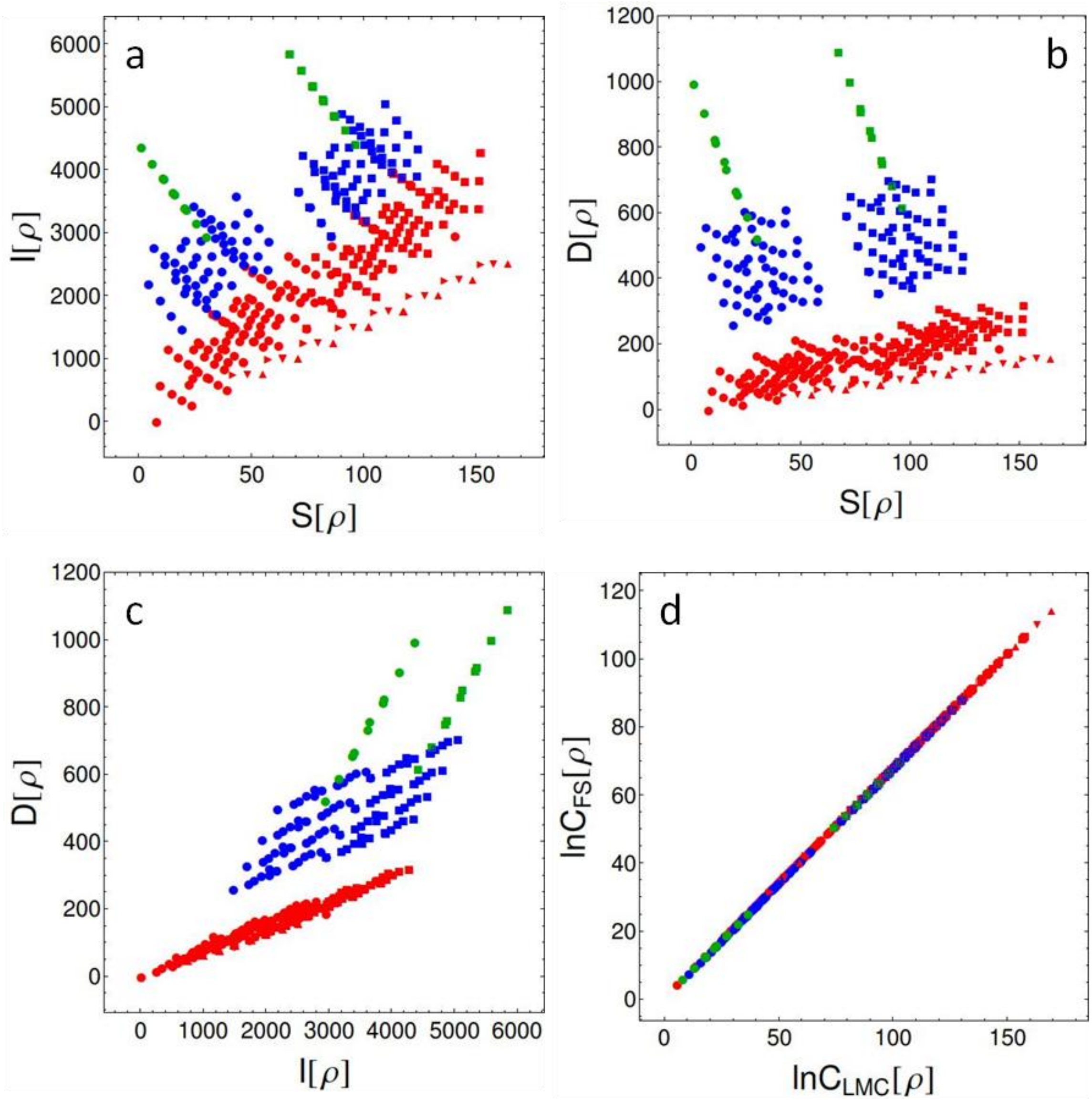

Fig. V-6. The IT measure planes (panels a-c) and the plane between logarithms of $C_{\mathrm{LMC}}[\rho]$ and $C_{\mathrm{FS}}[\rho]$ (panel d). All panels are in the spinor density representation.

Table V-5. The Shannon entropy, the Fisher information, the Onicescu information, the electron-electron repulsion energy, the electron-nucleus attraction energy, for X–Y molecules with $X,Y=\{\text{-SiH}_3\text{, -PH}_2\text{, -SH, -Cl}\}$ and for hexane isomers. All data in a.u.

	$S[\rho]$	$I[\rho]$	$D[\rho]$	$E_{\text{ee}}=E_{\text{es}}+E_{\text{xc}}$	E_{en}
X–Y molecule					
SiH_3–SiH_3	30.18	2939.61	522.46	312.99	-1566.62
SiH_3–PH_2	25.57	3161.13	589.89	331.54	-1687.13
PH_2–PH_2	21.05	3382.89	657.33	349.09	-1805.64
SiH_3–SH	20.47	3399.77	668.43	351.50	-1818.38
PH_2–SH	16.09	3621.68	735.88	368.04	-1934.82
SiH_3–Cl	15.08	3656.17	759.16	372.28	-1959.33
SH–SH	11.23	3860.45	814.47	386.22	-2062.51
PH_2–Cl	10.82	3878.24	826.64	388.13	-2074.39
SH–Cl	6.08	4117.09	905.20	405.63	-2200.76
Cl–Cl	0.95	4373.62	995.95	425.21	-2339.49
hexane isomers (C_6H_{14})					
3-methylpentane	102.30	1517.54	94.71	318.53	-1034.03
2,3-dimethylbutane	102.24	1517.21	94.72	327.93	-1052.83
isohexane	102.19	1517.07	94.71	330.62	-1058.19
2,2-dimethylbutane	102.15	1516.78	94.72	336.82	-1070.60
hexane	102.14	1516.64	94.72	340.17	-1077.32

Let us consider separately the isolelectronic systems, taking as an example the X–Y-III set and the hexane isomers, Table V-5. The FI value increases with increasing sum of Z_A^2 (see the *K*-shell model from Eq.(5-20)), in particular, hexane isomers have a lower FI value than the X–Y-III set molecules. The sharpness, concentration or delocalization of the electronic cloud is measured by the SE as well by the Fisher information (FI). In the case of the X–Y-III set, entropy decreases significantly with decreasing number of hydrogens (decreasing number of bonds). In the case of the isosters, the entropy is greater for the system with the higher symmetry, eg. diphospane and silanethiol, 3-methylpentane and hexane. The results for the entropy in the density representation are consistent with the interpretation that increasing of the electron-nucleus attraction energy $|E_{\text{en}}|$ (reflecting the "structure" addition to the distribution) is thereby lowering the entropy, whereas decreasing of the electron-electron repulsion energy would raise the entropy[128] (in Table V-5, the energy components for the X–Y-III set and hexane isomers are shown). Comparing the hexane data (Table V-5), the OI is the less effective tool to conformers distinction. It is worth mentioning that the discussed regularities are for the systems with the same electron number so these conclusions are also valid in the shape representation. More discussion about the group effect on the IT can be find elsewhere. [170]

Unfortunately, the extension of the definition of complexities used here to the spinor-density dependent functionals is inconvenient because it yields complexity values as very large numbers. Therefore logarithms of them are convenient for analysis. In Fig. V-6.d, an excellent linear relation between the logarithms of both complexities is observed. However, this linearity is due to the dominant role of the SE dependent factors of both complexities. When rewritten as $S[\rho]=a\ln(I[\rho])+b\ln(D[\rho])+c$, it is satisfied with very low accuracy.

We can conclude these two parts that the shape function contains less information than the spinor density and this can yield wrong conclusion when the information measure is used for the system with different electron numbers. Secondly, both complexities used in this work do not give chemically reasonable descriptions.

V.D The IT measures and reactivity indices

In this part, we would like to find some correlation between the IT measures and the chemical reactivity indices. When the transferability concept is valid for the IT measures, the correlated properties should also possess such properties as a desired condition for obtaining a good correlation, e.g.

$$\begin{aligned}R[\rho_{d-XY}]&=aF[\rho_{d-XY}]+b=aF[\rho_{XY}]+b+a\left(F[\rho_{d-XY}]-F[\rho_{XY}]\right)\\&=R[\rho_{XY}]+aF\left[\rho_{d-\mathrm{C_6H_4}}\right]=R[\rho_{XY}]+R\left[\rho_{d-\mathrm{C_6H_4}}\right]\end{aligned}\qquad(5\text{-}30)$$

where $R[\rho]$ is a correlated property, $d-\mathrm{C_6H_4}$ is the $\mathrm{C_6H_4}$ group lacking hydrogen atoms at *d* position.

The chemical potential and the chemical hardness are defined in terms of the electronic energy for neutral system and its ionic forms, Eq.(5-15). Unfortunately, there is no electronic energy transferability and consequently such indices as the chemical potential and the chemical hardness do not show significant correlation with the IT measures. The lack of transferability for the chemical potential and the chemical hardness is illustrated in Fig. V-7.a-

b. Only a general decreasing trend can be observed between the hardness and the SE or the FI. In the shape representation these relations are less visible. The observed trends between hardness and complexities are rather inconsistent. In the C_{LMC} complexity case, there are no direct relations, e.g. the molecules with non-hydrogen atoms from II row have very close complexity but wide hardness range Fig. V-7.c). The hardness vs. C_{FS} complexity shows a general decreasing trend (Fig. V-7.d).

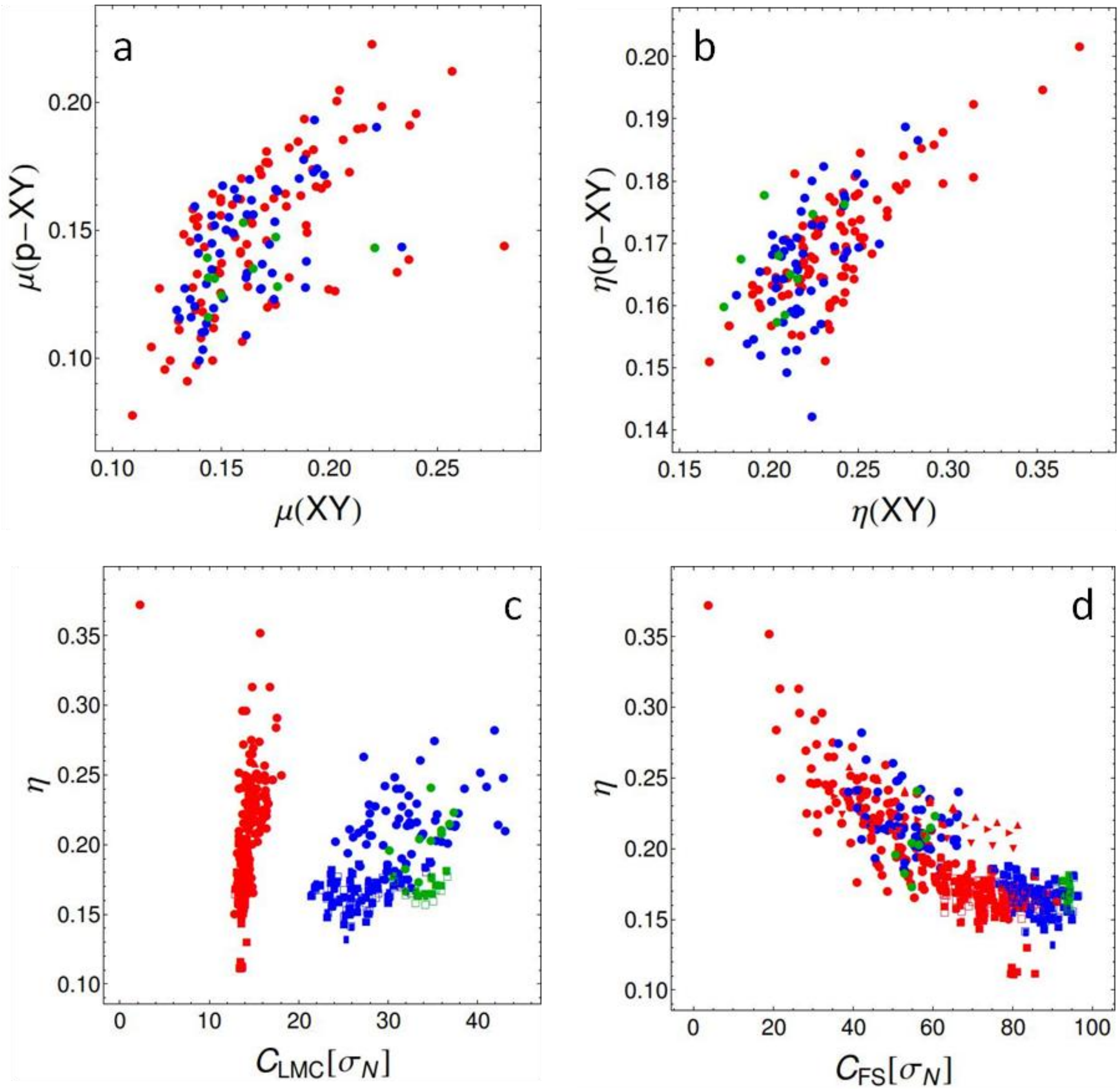

Fig. V-7. The lack of the group transferability for the chemical potential and hardness on the $p-\mathrm{C_6H_6}$ example in the spinor density representation (panels a-b). The chemical hardness vs. complexities in the shape representation (panels c-d).

In contrary to the electronic energy, the total energy shows the group transferability, thus the analysis of IT measures versus the total energy and its components was performed. Only the FI shows a linear behavior with the total energy, the kinetic energy and the nucleus-electron attraction energy with correlation coefficients 0. 967, 0.967 and 0.947, respectively. This behavior can be explained by recalling that the expectation value of the kinetic energy of non-interacting electrons of the Kohn-Sham problem is proportional to the Fisher information. When the virial theorem is satisfied with reasonable accuracy, the linear relation between the FI and the total energy is no surprise. The main contribution to the FI comes from each region close to nucleus, the same is true in the case of the nucleus-electron attraction energy.

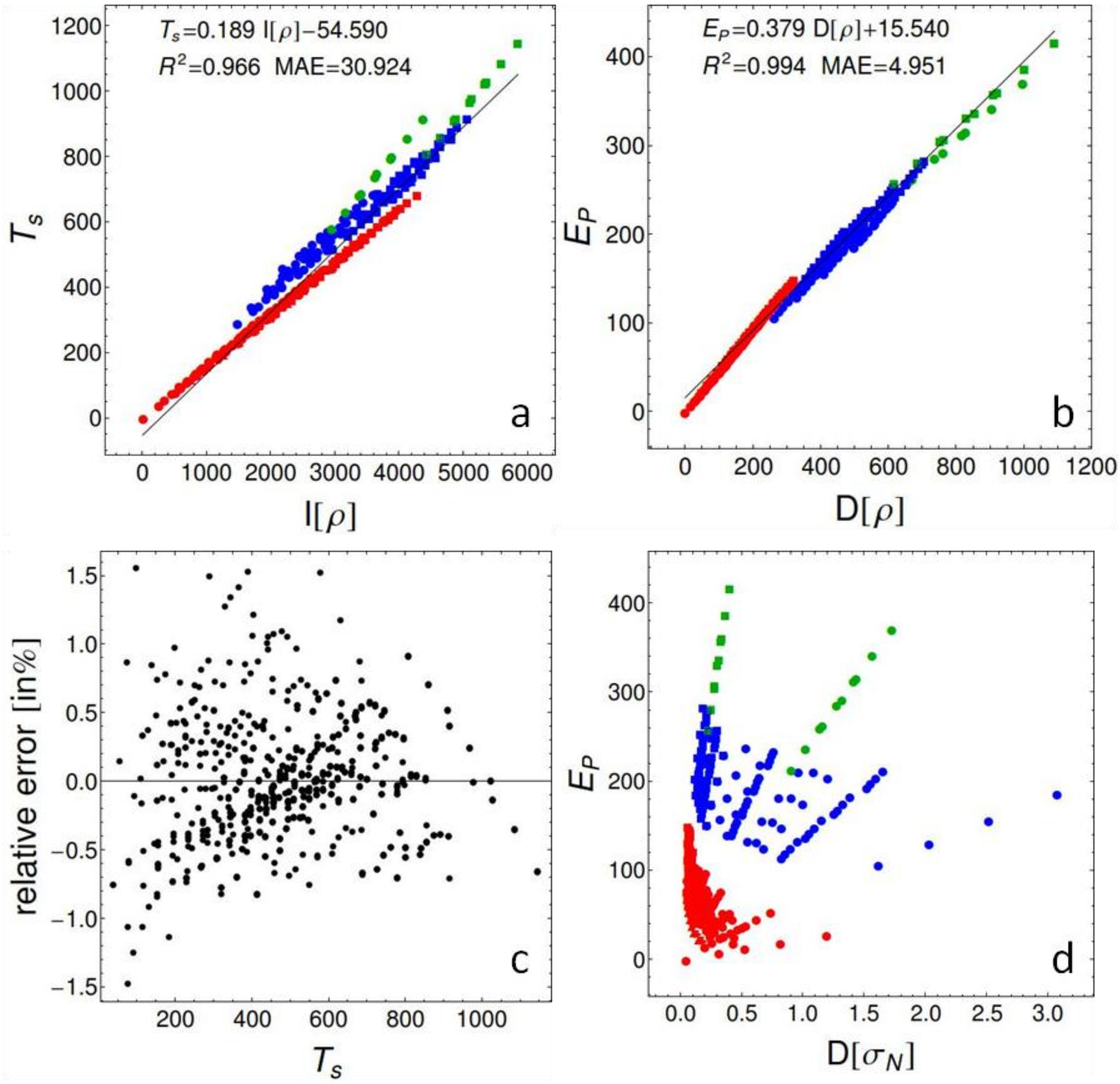

Fig. V-8. The correlation between the kinetic energy of non-interacting electrons and the Fisher information (T_s vs. $I[\rho]$, panel a); between the Pauli energy and the Onicescu information (E_P vs. $D[\rho]$, panel b); the relative error $\left(\tilde{T}_s - T_s\right)/T_s$ vs. T_s, of $\tilde{T}_s$ fitting, Eq.(5-31) (panel c); between the Pauli energy and the Onicescu information in the shape representation (E_P vs. $D[\sigma_N]$, panel d).

In addition to the linear behavior of T_{s} vs. $I[\rho]$ characterized by $R^2 = 0.966$ (Fig. V-8a), surprisingly, we observe a similar linear behavior in the case of the Pauli energy vs. the Onicescu information (disequilibrium) with $R^2 = 0.994$ (see Fig. V-8.b). These two linearities combined with the Pauli energy definition, Eq.(5-13), suggest T_{s} to be linear in $I[\rho]$ and $D[\rho]$. A multi-linear regression (through the origin) between the kinetic energy and the FI and the OI was calculated

$$\tilde{T}_{\mathrm{s}} = 0.1347\, I[\rho] + 0.3367\, D[\rho],\ \ \mathrm{R}^2 = 1.0000,\ \ \mathrm{MAE}{=}1.5745. \qquad (5\text{-}31)$$

The plot of the relative error, $\left(\tilde{T}_{\mathrm{s}} - T_{\mathrm{s}}\right)/T_{\mathrm{s}}$ vs. T_{s} of this fitting is presented in Fig. V-8.c. It is very small (except the hydrogen molecule error, equal -10%, not displayed). This fitting, Eq.(5-31), due to the relation $T_{\mathrm{W}} = I/8$ can be rewritten for the Pauli energy as $\tilde{E}_{\mathrm{P}} = 0.0097\, I[\rho] + 0.3367\, D[\rho]$. Since the virial theorem is approximately satisfied in DFT, Eq.(5-31) provides a model for the total energy $W \approx -T_{\mathrm{s}}$ with quite good accuracy. It is noteworthy that these linear statistical models disappear in the shape representation, e.g. compare E_{P} vs. $D[\sigma_{\mathrm{N}}]$, Fig. V-8.d, with E_{P} vs. $D[\rho]$, Fig. V-8.b.

Another interesting regression found, is the atomization energy ΔW vs. the atomization entropy $\Delta S[\rho]$, Fig. V-9.a. The atomization energy is the energy required to split a molecule into its constituent atoms, in other words to broke all the bonds, Eq.(5-16). The atomization entropy is defined as

$$\Delta S[\rho] = \sum_{A \in M} S\left[\rho_A^0\right] - S\left[\rho_M\right] \qquad (5\text{-}32)$$

where ρ_M is the molecule density and ρ_A^0 is the electron density computed for the isolated atom *A*. This ΔS is a modification of the SE due to the atomization (or the negative value of the SE change due to the formation of chemical bonds). The trend observed in Fig. V-9.a is in line with the interpretation of the entropy change during the bond formation from the promolecule.[171] As a consequence of the observed linear relation between the sum of atomic entropies and the molecular entropy, Fig. V-9.b, a linear relation between the atomization energy and the molecular entropy is observed, Fig. V-9.c. A positive slope indicates that the increase of molecular SE is accompanied by the increase of the energetical stability ΔW. However, in the case of the constitutional isomers, e.g. the benzene derivatives for a given X and Y substituents, these observations are deemed to be inconsistent. When the sum of atomic quantities is the same for constitutional isomers, from Fig. V-9.a and Eq.(5-32) follows $S_A > S_B \Rightarrow \Delta S_A < \Delta S_B \Rightarrow \Delta W_A < \Delta W_B$. In other words, if the entropy of molecule *A* is higher than the entropy of molecule *B*, then molecule *B* is more stable than molecule *A*, in contradiction with the observed relation in Fig. V-9.c. This indicates that the linear relation ΔW vs. $\Delta \mathrm{S}[\rho]$ is only a theoretical idealization of a general trend and this relation should not be used to compare molecules with small entropy difference.

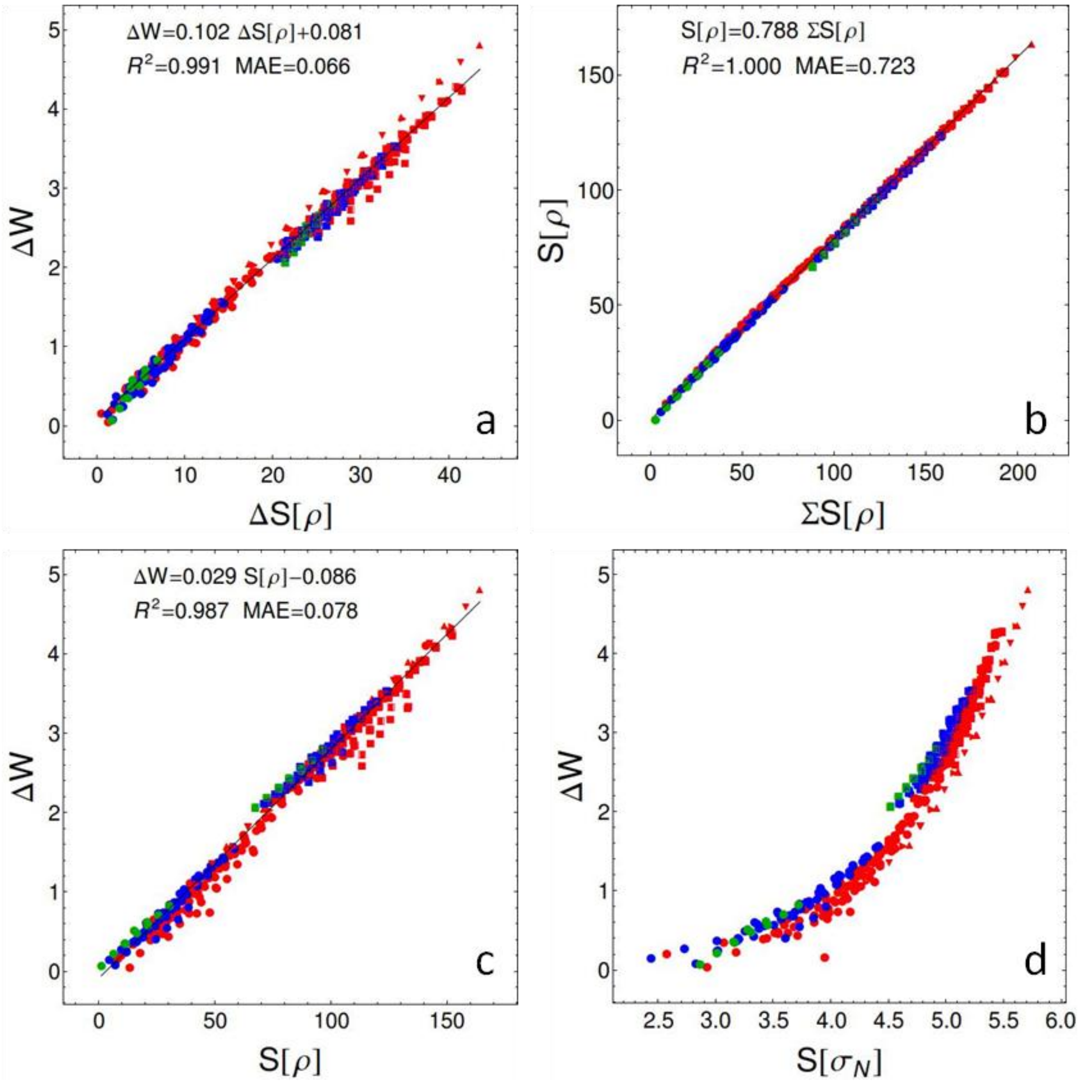

Fig. V-9. The correlation between: the atomization energy and the atomization entropy (ΔW vs. $\Delta S[\rho]$, panel a); the molecular entropy and the sum of atomic entropies ($S[\rho]$ vs. $\sum S[\rho]$, panel b); the atomization energy and the molecular entropy (ΔW vs. $S[\rho]$, panel c); the atomization energy and the molecular entropy in the shape representation (ΔW vs. $S[\sigma_{\mathrm{N}}]$, panel d).

V.E Conclusions

The analysis of the information and complexity measures as the tool in the investigation of the chemical reactivity and similarity has been done for the density representation in the spin-position space and the shape representation in the position space. The used set of molecules was enough large and diverse to improve the previous understanding of IT measures in chemical applications and to generalize the obtained conclusions. The concepts of the transferability and addititivity of atoms or functional groups were used as "checkpoints" in analysis of the obtained results. It was shown that in the shape representation, observed trends are in contradiction with chemical intuition, e.g. the FI value for alkynes > alkenes > alkanes irrespective of the number of carbon atoms they contain. The perfect linear relation between the XY molecules and benzene with the X and Y substituents observed in the spinor-density representation is devastated in the shape representation. In the case of the hydrocarbons investigated in this work, the entropy trend is opposite to the disequilibrium trend for a given carbon number. When the OI is considered as a finer measure of dispersion distribution than the SE, the previous observation is in contradiction with the meaning of both measures. The FI results are inconsistent with the requirements that the steric effect needs to be extensive in size (the FI for alkenes are almost constant with increasing size in the shape representation). The obtained values for the C_{LMC} and the C_{FS} complexities were mutually inconsistent, e.g. (i) in the case of A3 set, with increasing hydrocarbon size C_{LMC} is decreasing and while C_{FS} increases; (ii) the complexity trend observed for the changing substituents of the X–Y-III set is opposite to trend observed for the benzene derivatives, the d-X–Y-III set (the lowest complexity in the X–Y-III set is for chlorine molecule and the highest for disilane, while in the case of the d-X–Y-III set, the lowest complexity is for dichlorobenzene and the highest for disilylbenzene). In general, we can conclude that the shape function contains less information than the spinor density. Use of the shape function can lead to wrong conclusions when the information measure is analyzed for the systems with different electron numbers.

In contrary to the results from the shape representation, the measures S , I , D in the spinor density representation show the transferability and additivity. The group transferability is well illustrated on the example of the X–Y molecules and their benzene derivatives. As it was shown, a perfect linear relation exists between the SE for the X–Y molecules and the SE for their derivatives. Another example is the methylene group transferability presented on the alkane-alkene-alkyne set. Analysis of the information planes between the three IT based measures has shown that the $S-I$ plane provides "richer" information about the pattern, organization, similarity of used molecules than $S-D$ and $I-D$ planes. In the spinor-density representation, the C_{LMC} complexity and the C_{FS} complexity are inconvenient quantities due to their large values. They do not provide a chemically reasonable description of the molecular complexity. Unfortunately, no general relations between the IT measures and the chemical reactivity indices (like the chemical potential and the chemical hardness) are observed.

Surprisingly, in addition to the linear behavior of T_{s} vs. $I[\rho]$, a similar linear behavior in the case of the Pauli energy vs. the Onicescu information (disequilibrium) is observed. Based on these relations, the highly accurate linear relation is noted between the kinetic energy and the FI and the OI measures. Another interesting regression was found between the atomization energy and the atomization entropy. To the best of our knowledge, these linear behaviors have not been observed before. Finally, the atomic additivity in the spinor density representation proves the advantages of this representation over the shape representations. Of course the shape representation is still useful in work with systems which have the same electron number, e.g. for the reaction path analysis.

The results obtained in this work reveal that the IT measures can be used in the chemical reactivity investigation, but only as the source of the information about the pattern, organization, similarity of molecules, while not as direct indicators of their reactivity. In efforts to improve our understanding of the IT measures in the chemical applications, it seems necessary to take into account the substituent group influence on the IT measures and their relations with the reactivity change of substituted molecule.

Chapter VI. Analysis of Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons

The association of human cancer with the exposure to polycyclic aromatic hydrocarbons (PAHs) dates back to the second half of the 18th century. PAHs are organic chemical compounds, found usually as complex mixtures, contain two or more aromatic rings fused together in a linear (*cata*-condensed) or angular (*peri*-condensed) configuration. The PAHs are formed from all sorts of incomplete combustion of organic materials and hence are persistent and abundant throughout the environment so as to have negative or positive consequences to living things. The diesel engine exhaust, industry and the burning of wood (anthropogenic sources) are the main sources of PAHs. They are widely used in synthesis of dyes, pesticides, medicine, asphalt, etc. Due to their abundance, persistence and toxicity nature, PAHs have been widely studied at different levels in the field of environmental toxicology. Their mode of toxicity can cause tumor and may also lead to the development of cancer or other cellular problems.[172] The mechanism of carcinogenicity is a complex process and is a challenging task to find a universal theoretical model for the characterization and prediction of carcinogenic effect of molecules. Although a number of successful researches have been done on the carcinogenicity of PAHs[173-179] so far either the amount of data used or the accuracy achieved were not so attractive. The types of descriptors as well as the amount of species (data set) used are the key factors that determine the stability and quality of any model

In this chapter, the carcinogenic activity (CA) of PAHs is investigated in two ways. First, the carcinogenicity database was constructed. Peripheral regions (using graph theory principle) and E-Dragon descriptors along with the support vector machine (SVM) based model of carcinogenicity is applied. The E-Dragon software is an application for the calculation of molecular descriptors. These descriptors can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and high throughput screening of molecule databases.[1,2] The SVM is a supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. The base of SVM is presented in Appendix B.

Secondly, the alchemical deprotonation energy of one or two protons is used in the qualitative investigation of the CA of PAHs.

At first the carcinogenic effects and mechanisms of PAHs are shortly revisited. Then, the peripheral regions of PAHs are defined. Using the selective molecular descriptors included in the E-Dragon software and the descriptors related to the peripheral regions (peripheral and E-Dragon descriptors), SVM model is developed. Finally, the alchemical deprotonation energy is tested for exploring (qualitatively) the reactive sites of PAHs.

VI.A Carcinogenicity and mechanism of action

PAHs are stable and need metabolic activation to exert their carcinogenic activity by means of biotransformation to chemically reactive intermediates such as epoxide intermediates, dihydrodiols, phenols, quinones, and so on.[172] Our body struggles to detoxificate any xenobiotic (external molecules or drugs) into water soluble products that could be removed from the body via excretion pathways. However, the highly unstable carbocation intermediates can covalently bind to cellular nucleophiles (such as DNA) that leads to tumor. When it is malignant (primary growth can generate several secondary growths thus invading or spread all vital parts of our body), a tumor becomes cancer (though not all tumors are cancerous or vise versa).

There are two main metabolic activation pathways of PAHs in the initiation (activation) of cancer[172] namely the one-electron oxidation (radical reaction) and two-electron oxidation (monooxygenation). The former one occurs due to relatively high charge localization cation and ionization potentials below ca. 7.35 eV (easy metabolic removal of electron).[180,181] The two-electron oxidation takes place due to formation of diol-epoxides (carbocation intermediate) to form trans-dihydrodiols [172,182]. Those unstable carbocations can react with DNA to form DNA adduct and aprunic site and becomes cancer. For instance, binding of benzo[a]pyrene to mouse skin DNA occurs predominantly at L -region (C-6), the position of highest charge localization in its radical cation[183] Cyctochrome P450[184] and microsomal epoxide hydrolase (mEH)[185] are the main enzymes to undergo the two-electron oxidation pathway (see Ref.[172] for details).

Many attempts have been made to devise a system suitable for the accurate comparison of the potency of carcinogens. The carcinogenicity index, called Iball index, was first suggested by John Iball[186]. It is defined as the percentage number of tumor incidence to the mean latent period (in days) multiplied by 100. The mean latent period is the ratio of the number of animals with tumor to the number of animals alive when the first tumor appears. The logarithm of Iball index along with experimental CA or carcinogenic potency (CP) of PAHs are given in the SID. The log (Iball) is given by assigning numerical values to define relative biological activity such that inactive PAHs have zero value (activity of unity) [187]. Note that the CA or CP is based on the maximum possible carcinogenic effect in any animal or human part of the body (a molecule can be non-carcinogenic when tested in skin but carcinogenic in tissue, so it is considered carcinogenic).

Another quantitative measure of biological activity (can be mutagenic, carcinogenic or other toxicity nature) is the amount of hydrophobicity, measured by the logarithm of the octanol water partition coefficient P (ratio of $C_{\mathrm{m,octanol}}$ to $C_{\mathrm{m,water}}$) with C_{m} being the molar concentration of the solute in octanol or water at a specified temperature)[188]. As human body is made up of water and lipids (hydrophobic), knowing the distribution ratio of the chemicals to the octanol, one can estimate the biological accumulations in our body. The relative hydrophobicity of the chemicals determines the amount of carcinogenicity, this in turn depends on the presence of substituents on the L and K -regions of the carcinogen[187]. According to this, the carcinogenic potency has an increased behavior up to log P of 6.14 and then shows a decreasing trend for log P above 6.14.

Because of the complexity of chemical carcinogenesis, linear correlations of carcinogenic potency with a single theoretical concept are a very crude first step. In the case of PAHs, the influence of distinct molecular regions is well established (see next section for regions definitions). The experimental evidence for the metabolism via epoxide, dihydro diol, and dihydrodiol epoxide to a bay-region carbocation reacting with DNA clearly points to the pertinent reactivity centers. In addition, molecules containing reactive L regions, such as polyacenes, are known for not being carcinogenic. Such chemical structure features or properties that alter physicochemical properties of chemicals, molecular descriptors, are used to develop the quantitative structure activity relationships (QSARs). It is since 1955 Pullman's theory [175] and later Lehr and Jerina[176] describe a theoretical basis for the relationship between peripheral regions (such as the B -region theory) and biological activity (carcinogenicity for example) of PAHs.

VI.B Peripheral regions of polycyclic aromatic hydrocarbons

Peripheral regions of PAHs are a type of topological structures that describe external (exterior) part of the aromatic ring only. These regions were categorized into two main groups. The first classification base on the number of consecutive sp^2 and/or sp^3 carbons containing hydrogen (carbon-hydrogen sp^2 or sp^3-hybridized) within the aromatic ring (see Fig.VI-1). The L -region has one number of sp^2 or sp^3 carbons containing one or two hydrogen/s as part of the peripheral aromatic ring system while K , N , M and O regions are defined with 2, 3, 4 and 5 number of consecutive hydrogen containing sp^2 and/or sp^3 hybridized carbons respectively. This group will be called the LKNMO group regions.

The second classification is based on the number of consecutive non-hydrogen sp^2 (hybridized) carbons within the aromatic ring. The following regions can be distinguished: the bay region (B) with 2 consecutive non-hydrogen sp^2 carbons, the fjord region (F) with 3 sp^2 carbons, the harbor (H) and canyon (C) regions with 4 and 5 sp^2 carbons, respectively. These groups will be called BFHC group regions. Note, any substituents within the same ring cannot form any of the BFHC group (lacking of the fusion carbon), for example, the region between the two methyl (-CH_3) in Fig.VI-1 is not B -region.

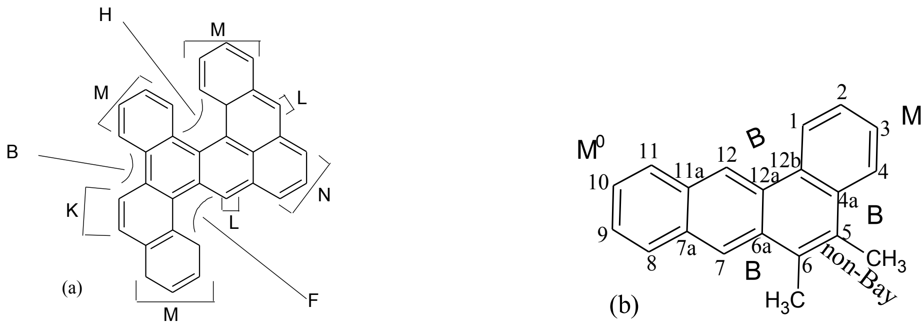

Fig.VI-1. (a) Labeling of peripheral regions (B ,F ,H ,L ,K ,N , and M) of PAHs. (b) Nomenclature rule of numbering carbons . M -region that is not neighbor to the BFHC group is denoted by M^0. The CH3 is a methyl substituent. Any substituents within the same ring cannot form any of the BFHC group (lacking of the fusion carbon).

VI.C Prediction of chemical carcinogenicity by SVM approach

VI.C.1 SVM classification models

In this subchapter, I report a successful application of machine learning approaches, namely SVM approach was evaluated for predicting carcinogenicity of PAHs using molecular descriptors. A set of 302 compounds was used to estimate the accuracies of three models. Out of 302 PAHs 269 molecules have assigned the experimental CA or CP (33 PAHs lack carcinogenic activity information) while 117 out of 302 molecules have assigned the logarithm of Iball index of carcinogenicity and 131 out of 302 PAHs have log *P* (see SID).

The three SVM classification models were proposed and tested

-model I. This model is based on the peripheral regions. According to different scientific evidences of carcinogenic activity[175-177], the regions belonging to the $BFHC$ group are activating (promoting) ones while within the $LKNMO$ group, K and M regions are activating whereas L and N regions are deactivating ones. Hence, we can try the peripheral index D defined by

$$D = |P - Q|. \qquad (6\text{-}1)$$

to be used as carcinogenic activity descriptor. Here P is the total number of regions belonging to $\{K, M, B, F, H, C, O\}$ and Q is the total number of regions belonging to $\{N, L, M^0\}$. Note that M^o is a type of M-region that is not neighbor to the B, F, H, C, or O-regions as shown in Fig.VI-1(b). Values of all the peripheral regions for 302 PAHs are given in SID.

-model II. This model uses 11 selected E-Dragon molecular descriptors [2,189] (IC5, SIC5, CIC5, Mor29u , Mor31u, Mor29e, Mor31e, E2m, R6u, C-024, C-025). The first three descriptors i.e. IC5 (Information content index), SIC5 (Structural information content) and CIC5 (Complementary information content), are information indices with neighborhood symmetry of 5-order. The Mor31u (3D-MoRSE - signal 31 / unweighted), Mor31e (3D-MoRSE - signal 31 / weighted by atomic Sanderson electronegativities); 3D-MoRSE - signal 29 / weighted by atomic Sanderson electronegativities (Mor29e) and Mor29u (3D-MoRSE - signal 29 / unweighted) are descriptors calculated by summing atom weights viewed by a different angular scattering function[190]. The E2m (2nd component accessibility directional WHIM index / weighted by atomic masses) is a type of WHIM descriptors obtained as statistical indices of the atoms projected onto the 3 principal components obtained from weighted covariance matrices of the atomic coordinates. The R6u (R autocorrelation of lag 6 / unweighted) is calculated from the leverage matrix obtained by the centered atomic coordinates[191]. The C-024 and C-025 are the atom-centred fragments for R--CH--R and R--CR--R (R is fragment atom or molecule), respectively. For details about molecular descriptors, I consult the readers to look the *Handbook of molecular descriptors*[189] and Ref. [192].

-model III. uses 12 descriptors selected from both peripheral and E-Dragon descriptors (L , K , M , M^0, D , IC5, SIC5, CIC5, Mor31u, E2m, C-024, C-025). The target instance is the carcinogenic activity (CA). The random forest algorithm[193] and the generalized cross-validation within the earth package[194] are used to select the most significant predictor variables.

The SVM approach can also be applied for regression estimation. The SVM regression uses the log (Iball index) and log (*P*) values (as target instances) with the selected peripheral and E-Dragon descriptors. The M , B ,CIC4[2,192] and EEig09 [195] descriptors are used to develop predictive model for log (Iball index). The CIC4, abbreviated for the complementary information content neighborhood of 4th order symmetry, is a group of information indices that is related to Shannon (information) entropy[196]. The EEig09, abbreviated for the eigenvalue 09 from edge adjacent matrix weighted by resonance integrals, is a group of edge adjacency indices [2,189]. On other hand, to model log (*P*), the D (Eq. 6-1)) as well as the DECC (average eccentric) [189,197] and piPC04[2] (molecular multiple path count of order 10) from E-dragon descriptors, are used.

The support vector classification and regression implementations of the Kernlab [198] package of R version 3.0.2 (along with the Rattle package)[199] were used to construct the model. The accuracy of the (K)SVM models are compared with other machine learning methods such as decision tree (RPART) [200], random forest (RF)[193], neural networks (NNET) [201] and adaptive boosting (ADA) [202].

VI.C.2 Models validation. Results and discussion

VI.C.2.a Confusion matrix

The central part of performance measure is relayed on the values within the so-called confusion matrix[203]. A confusion matrix is a cross-tabulation of the predicted and observed classes used as a measure of performance of classification model and has size of *n* by *n* (*n* being the number of different classes). The most commonly used, however, is when *n* = 2 for positive and negative classes. For two-class problem, four different possible outcomes of a single prediction are displayed. These are represented by columns (predicted values) and rows (actual values) as shown in Table VI-1. The number of correct predictions that an instance is negative is denoted by TN or positive (TP) while the number of incorrect predictions that an instance is negative (FN) or positive (FP). Confusion matrix gives very much information about the classifier that shows how many of the actual data points are predicted correctly.

Table VI-1. Elements of confusion matrix

		Predicted value	
		Negative class (non-carcinogens)	Positive class (carcinogens)
Actual value (observed)	Negative class (non-carcinogens)	True negative (TN)	False positive (FP)
	Positive class (carcinogens)	False negative(FN)	True positive (TP)

The receiver operator characteristic (ROC) curve is the common measure of performance of SVM classification. The ROC curve is the plot of true positive rate (TPR), also called recall or sensitivity

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{6-2}$$

versus the false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{6-3}$$

The ROC curve of the point (1,1) indicates a perfect classifier. A random model without discriminative power has an area under the ROC curve (AUC) value of 0.5. The total accuracy (the ratio of the total number of correctly predicted carcinogens and noncarcinogens to the total number of molecules) is

$$\mathrm{Accuracy} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}. \tag{6-4}$$

The specificity, also known as true negative rate, is the percentage of correctly classified noncarcinogens among the total number of noncarcinogens. It is defined by

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{6-5}$$

Cross-validation (CV) is a standard technique for adjusting hyperparameters in developing a predictive model. One example of cross-validation is the *K*-fold cross-validation where the training data splits into *K* roughly equal-sized parts. Using the k^{th} part (k = 1, 2, …,K), the model is fitted to the other *K*-1 parts of the data and then calculated the prediction error of each k^{th} part and combined the *K* estimates of prediction error. The optimal parameters such as the cost *C* (general penalizing parameter) and gamma (the specific kernel parameter) are obtained by grid search along with cross-validation. This search performs the CV for each element of the cross-product and selects the one with the best mean performance (lowest CV error). CV is one way to determine realistic standard errors and controls over-fitting (a situation when the model requires more information than the data can provide or introduction of more parameters).

A special case of k-fold CV with k equals the number of instances (rows) in the data is called leaving-one-out (LOO)CV[204]. This is important for the reasons of estimating the bias of the excess error in prediction as well as gives better tuning parameter description.

VI.C.2.b Classification

Three SVM models namely model-I, model-II and model-III are proposed —depending on the type of descriptors used. The 80% training and 20% testing sets is found to be the best choice for model-I and Model-II, while model-III was performed using the 70% training and 30% testing set. Grid searches with 10-fold cross-validation and LOO CV were performed to find the optimal parameters. As shown in Table VI-2 the (optimal) cost parameter *C* for model-I, model-II and model-III is 15.1, 16, and 8, respectively. The polynomial (with degree 4), radial basis function (RBF) and Laplace function kernels are used for the respective models (see Appendix for kernels definition).

From the descriptors used to construct model-I, the L , K , M -regions (and somehow $\text{Specificity} = \text{TN}/(\text{TN}+\text{FP})$ and D) have high weighting factor (sum of coefficients of the support vectors) and this indicates that carcinogenic effects are more prone to the changes in these three regions than others. This model can predict with an accuracy of 86.24% (232 out of 269 PAHs). The optimal parameters are gamma = 1.01, *C* = 15.1 and d = 4. The first row in the confusion matrix of Table VI-2 for model-I indicates that 118 objects belong to the class of noncarcinogens, out of which 102 are correctly classified as noncarcinogens and 16 misclassified as carcinogens.

ROC curve (can be also defined by the sensitivity as a function of fall-out) provides a useful graphical assessment of the performance of a classifier by increasing rate of false positive rate errors that are tolerated to improve the true positive rate. The point (0,1) is the perfect classifier that enables to classify all positive (carcinogens) and negative (noncarcinogens) cases correctly (the false positive rate is zero while the true positive rate is 1). The area under the ROC curve (AUC) is equal to the probability that a true positive is scored greater than a true negative. Hence, a model having ROC curve (upper left hand corner) close to 1 is better model. We can look in Fig. VI-2(a) and (b) for mode-I and model-

II with AUC values 0.8874 and 0.9717, respectively while model-III (Fig. VI-3(a)) has value of 0.9902.

Table VI-2. Confusion matrix and overall performance measures of SVM classification models I, II and III. The confusion matrix is shown by the actual versus predicted values. The total accuracy, sensitivity, specificity and area under the ROC curve (AUC) are given for each model. A type of kernels, optimal parameters, errors and number of support vectors are also shown. Attributes count the number of descriptors plus for the target parameter.

		Predicted*					
Actual		Model-I		Model-II		Model-III	
		Non-carcinogen	Carcinogen	Non-carcinogen	Carcinogen	Non-carcinogen	Carcinogen
	Non-carcinogen	102	16	103	10	105	9
	Carcinogen	21	130	9	138	7	140
Accuracy		0.8624		0.9269		0.9386	
Sensitivity		0.8606		0.9387		0.9523	
Specificity		0.8644		0.9115		0.9210	
AUC		0.8874		0.9717		0.9902	
Kernel		Polynomial (d = 4)		RBF		Laplace function	
Cost (C)		15.1		16.0		8.0	
Gamma		1.01		0.0908		0.0910	
Training error		0.0976		0.0432		0.0219	
Cross-validation error		0.3987		0.2790		0.1973	
No. of support vectors		141		119		138	
Bias (*b*)		-0.0929		0.0387		0.0015	
Attributes		12		12		13	
Data examples**		269		260		261	

*The confusion matrix here is constructed only for those molecules whose experimental carcinogenic activities (CA) are known. **Different data sets are used due to some descriptor values are missing (see SID).

Comparisons of performances of these three SVM models are shown in Table VI-2. Model-III is the best model in all performance measures. Hence, apart from some comparison and characteristic features of model-I and model-II, detailed analysis and prediction of carcinogens/noncarcinogens will be done using model-III.

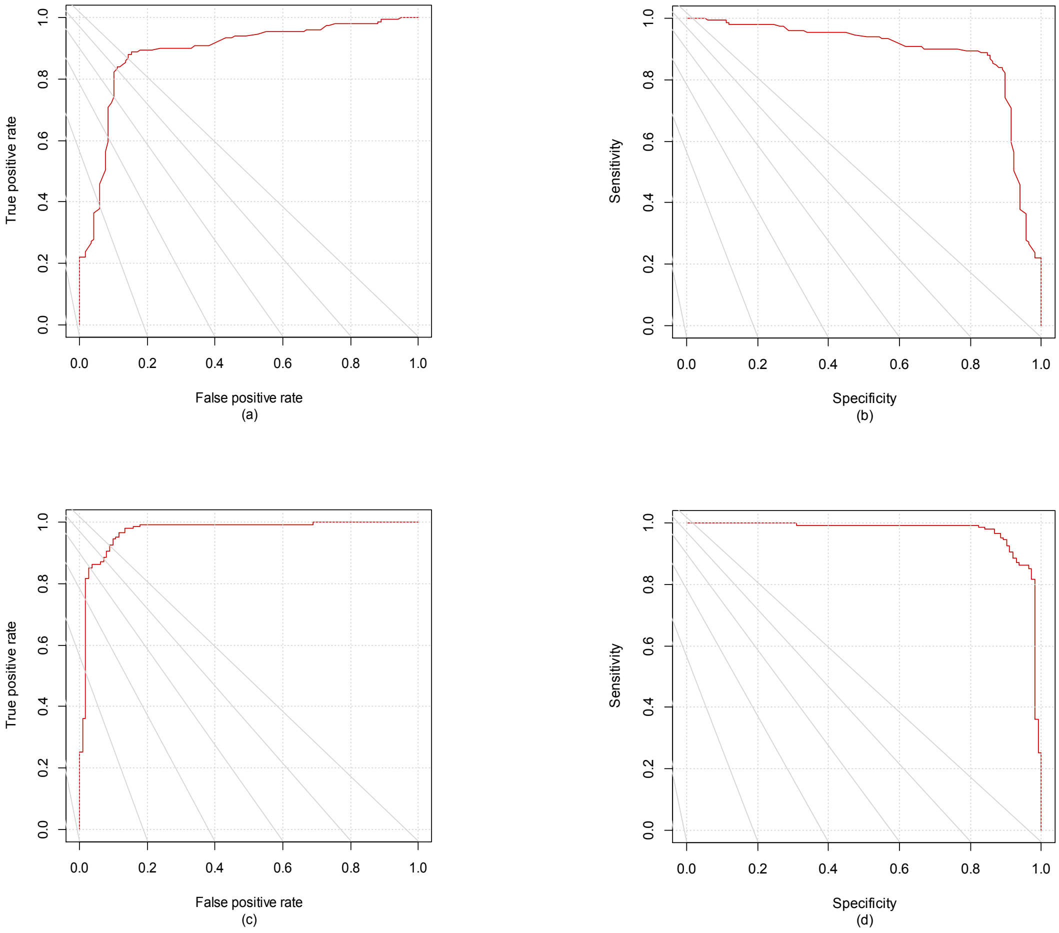

Fig. VI-2. (a) The ROC curve and (b) sensitivity vs. specificity for model-I. (c) ROC curve and (d) sensitivity vs. specificity for model-II

The selected E-Dragon descriptors are used to construct Model-II with more accurate results than model-I. This has accuracy = 92.69%, sensitivity = 93.87%, specificity = 91.15%. As shown in Fig. VI-2(c) and (d), the AUC (0.9710) and sensitivity versus specificity are relatively good, because the curve is close to the point (1,1) where all carcinogens (FN = 0) and all noncarcinogens (FP = 0) cases are obtained. This model can predict with an accuracy of 92.69% (241 out of 260 PAHs) —103 as noncarcinogens and 138 carcinogens (19 PAHs misclassified). The RBF kernel with optimal parameters gamma = 0.0908 and C = 16 are used.

SVM Model-III uses the Laplacian kernel and a grid search along with the 10-fold cross-validation (CV error of 0.1973) and OOV (CV error of 0.1964) are performed. As a result, the optimal parameters (hyperparameters) gamma = 0.0909 and cost (C) = 8 are obtained. There are 138 support vectors which determine the boundary decision between the carcinogenic and noncarcinogenic PAH molecules. The weighted sums of the coefficients of these support vectors and the bias form the decision function. The bias has value 0.0015, which indicates good model performance. Hence, we will have a decision function, in the form of Eq.(C-10) to determine which class it belongs.

Table VI-3. Confusion matrix and performance measures for training (70%), testing (30%) and overall sets of SVM model-III.

		Predicted*					
Actual		Training		Testing		Overall	
		Non-carcinogen	Carcinogen	Non-carcinogen	Carcinogen	Non-carcinogen	Carcinogen
	Non-carcinogen	77	4	28	5	105	9
	Carcinogen	0	101	7	39	7	140
Accuracy		0.9780		0.8481		0.9386	
Sensitivity		1.0000		0.8478		0.9523	
Specificity		0.9506		0.8484		0.9210	
AUC		0.9999		0.9374		0.9902	
Data examples		182		79		261	

*The confusion matrix here is constructed only for those molecules whose experimental carcinogenic activities (CA) are known (see SID).

The statistical performance of model-III is presented in Table VI-3. The dataset (261 PAH molecules) are splited into 70% (182 PAHs) as training set and 30% (79 PAHs) a testing set. Based on this, training set indicated an accuracy (in percentage) of 97.80%, high value of sensitivity (100%) , specificity (95.06%) and AUC 0.9999 (see Fig. VI-3). The training error is about 0.02197. For the test set, it gives accuracy equal to 84.81%, sensitivity (84.78%) and specificity (84.84%). The overall performance shows 93.86% accuracy, 95.23% sensitivity and 92.10% specificity. As a result, 245 out of 261 PAHs are correctly classified. Among these, 105 PAHs are classified as noncarcinogenic while 140 are carcinogenic. Of the remaining 16 misclassified PAHs, 7 of them are wrongly classified as noncarcinogenic and 9 are misclassified as carcinogenic PAHs (see SID). For example, dibenz[a,j]anthracene, dibenz[a,c]anthracene, 1,2,3,4,8,9-hexahydro dibenz[a,j]anthracene and tetracene are weak carcinogenic PAHs but are predicted as noncarcinogenic where as phenanthrene is classified as carcinogenic. Note that almost all those misclassified PAHs have very weak carcinogens.

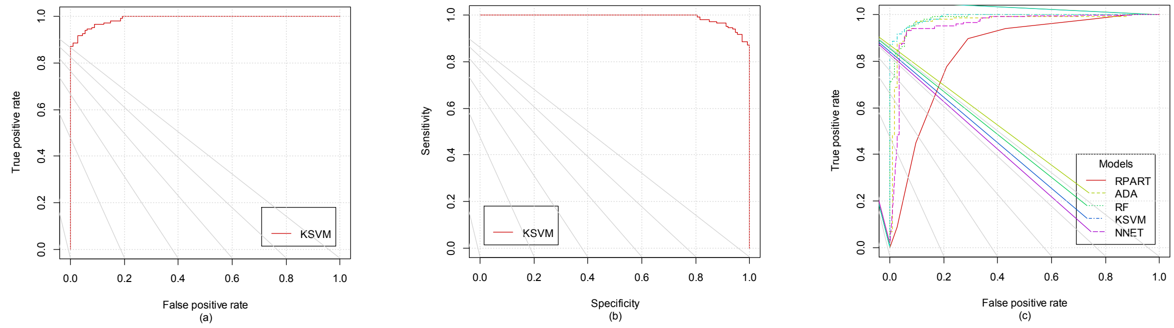

Fig. VI-3. (a) ROC curve and (b) sensitivity vs. specificity for model-III while (c) shows comparison of ROC curves of support vector machine (KSVM) with decision tree (RPART), adaptive boosting (ADA), random forest (RF) and neural network (NNET) models.

According to SVM model developed by O. Ivanciuc[179] PAH molecules such as 7-methyl benz[a]anthracene (**13**), Dibenz [a,h]anthracene (**4**), 5-methylchrysene (**32**), 6,8-dimethyl benz [a]anthracene (**142**) and 5-methylbenzo[c]phenanthrene (**162**) are all classified as inactive (noncarcinogens). From experimental evidences[187,205] all these molecules are carcinogens (active). Our model (model-III) predicts all these molecules as carcinogens (active). Moreover, the total accuracy of our model is much better than the general QSAR model for carcinogenicity[174].

The ROC curve and the sensitivity versus specificity plots for model-III are shown in Fig. VI-3. Both show good performance measures (see also Table VI-3) as both curves are close to 1. Fig. VI-3(c) is plot for comparing the SVM performance with other machine learning methods. The AUC for the RPART, ADA, RF, KSVM and NNET models are 0.8406, 0.9700, 0.9848, 0.9902, and 0.9558, respectively. As we can easily see, the KSVM experience high AUC value (followed by neural network and random forest).

VI.C.2.c Regression

Both log (Iball index) and log (*P*) SVM regressions use 80% training and 20% testing splitting scheme.The correlation between Iball index of carcinogenicity and selected descriptors M , B , CIC4 and EEig09 is modeled using the SVM regression.

According to descriptor selection, we observed that the M -region is more susceptible to carcinogenic potency than the B -region. The noncarcinogenic Dibenzo[b,k]chrysene (**173**) molecule, is good example, have 2, 2, 4 and 2 numbers of B , L , K and M^0-regions respectively and lacks M -region. Bold numbers in parenthesis refers to the serial number of the particular molecule in SID. Other examples include 1-methylbenz[a]anthracene (**7**), 2-methylbenz[a]anthracene (**8**), 3-methylbenz [a]anthracene (**9**), 4-methylbenz[a]anthracene

(**10**) while 9,10-dimethyl anthracene (**21**) and 1,2,3,4,8,9-hexahydro dibenz[a,j]anthracene (**22**) have M -region and posses slight carcinogenic activity. Note that molecule containing two or more consecutive M -regions does not have carcinogenic behavior. Examples of this are triphenylene molecules (**49**, **50** and **104**), phenanthrene (**24**) and 9-methylphenanthrene (**123**). On the other hand, due to presence of sterically-hindered F-region, molecules such as dibenz[a,c]fluorine (**71**) and dibenzo[a,c]phenanthrene (**106**) possess non-zero Iball index.

Some PAHs have none of the BFHC groups but possess carcinogenic activity. Azuleno [5,6,7- cd]phenalene (**302**) is highly potent which is probably due to the presence of 5 and 7-membered rings that affect the stability of the 6-membered rings. For the log (Iball index) as a target instance, the SVM regression uses the Laplace function kernel (Eq.(C-15)) with gamma equal to 0.3235. The other optimal parameters are *C* = 500, bias (*b*) = 0.7982 and epsilon = 0.1. The training and cross-validation errors are 0.00918 and 0.386087, respectively. Thus, the regression equation can be easily constructed (as in the form of Eq.(C-20)) using the coefficients of the support vectors of the M , B , CIC4 and EEig09r descriptors, the Laplace function kernel, gamma and bias.

The quality of this regression model can be seen in the predicted versus experimental log (Iball index) plot in Fig. VI-4(a). This gives a coefficient of determination (R-squared) value of 0.9475. The residual standard error is 0.1619 (on 115 degrees of freedom) with p-value less than 2e-16. Removing the outliners (red-star points) does not improve the correlation (even it makes a bit worse) because the number of slack variables increases dramatically (more points are outside the tube), as a result the training and cross-validation errors goes high as does the number of support vectors (or becomes over-fitted). Hence, these outliners are reasonable (5 points out of 117 molecules) to include in the model.

The SVM regression model for log (*P*) is developed using the D , DECC and piPC04. Unlike the Iball index (where M and B regions are more descriptive), this general biological activity log (*P*) depends on all peripheral regions (Eq.(6-1)). Two molecules namely Naphtho[1,2,3,4-def]chrysene(**36**), and 7,14-dibenzyldibenz [a,h]anthracene (**82**) are excluded from the regression model. These molecules have very high log (*P*) values.

In this regression, the Laplace function kernel with epsilon = 0.1, *C* = 100, gamma = 1.04077 and bias = -0.27928 are best chosen. There are 78 support vectors and errors 0.00866 (training) and 1.059655 (cross-validation) with objective function value -65.3784. Thus, the regression equation is constructed based on Eq.(C-20). The quality of the model can be shown in Fig. VI-4(b) where the predicted versus experimental log (*P*) values produce a correlation coefficient of 0.9597 with residual standard error 0.2269 on 127 degrees of freedom.

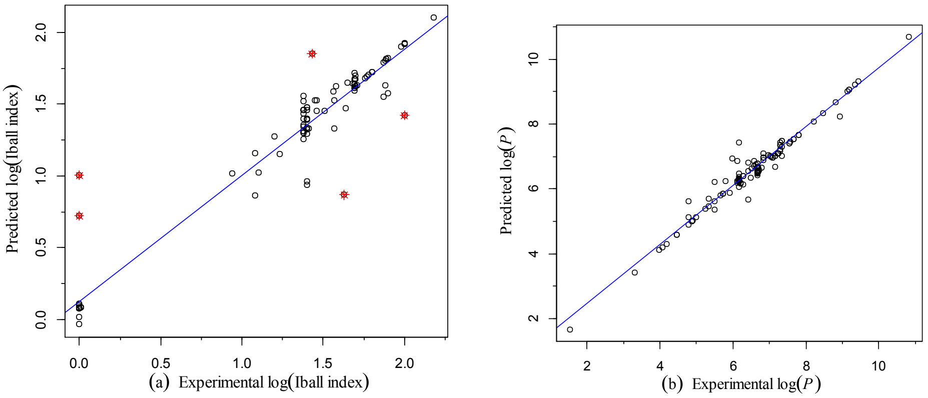

Fig. VI-4. Plots of (a) predicted versus experimental logarithmic (base 10) Iball index (removing the red star points does not improve the correlation, see text for detail) and (b) predicted versus experimental log (*P*).

Generally, we should bear in mind that the molecular descriptors used for SVM classification (where the experimental carcinogenic activity is used as a target instance) are not that much significant for regression (where the experimental Iball index and/or log (*P*) values are used). For example if we take the experimental Iball index of carcinogenicity, not all molecules that have zero log (Iball index) are carcinogens. This event is happening (as Iball himself pointed out) due to the fact that a number of animals may die soon after the first tumours are seen and before the majorities have appeared. Beside this the sensitivity of the experiment and other environmental conditions may produce different carcinogenic potency. For instance, both Naphtho[1,2,3,4-def]chrysene (**36**) and Dibenzo[a,e]pyrene (**39**) have zero log (Iball index) but they are potent carcinogens while the 1,2,3,4-tetramethylphenanthrene (**98**) with log (Iball index) value of 1.70 is carcinogenic inactive. Hence, the quality of these models should be evaluated in accordance to which experimental quantities or entities are used as a target values.

VI.D Qualitative study of PAHs using the alchemical derivatives

The alchemical indices will be tested here as PAH structure descriptors, not as the indices directly involved in the description of the generation of electrophilic metabolites of PAHs. The results presented below represent the tentative study on the usefulness of the alchemical indices in the application indirectly related to the nuclear charge changes.

The alchemical derivatives were calculated for the set of 12 PAHs at B3LYP/cc-pVDZ level. The considered PAHs are listed in Table VI-4 and presented in Fig. VI-5.

Table VI-4. Experimental Iball index of carcinogenicity of PAHs.

	hydrocarbons name	Iball index[206]
I	anthracene	0
II	benzo[a]anthracene	7
III	benzo[a]pentacene	0
IV	benzo[a]pyrene	72
V	benzo[e]pyrene	2
VI	chrysene	5
VII	dibenzo[a,e]pyrene	50
VIII	dibenz[a,j]anthracene	4
IX	dibenzo[e,l]pyrene	0
X	naphthalene	0
XI	naphtho[2,3-a]pyrene	27
XII	phenanthrene	~

For all PAHs the alchemical deprotonation energy for one and two side deprotonation were calculated according to the scheme presented in **IV.A** section. The one side deprotonation energy from Eq.**(4-4)** is

$$D_A^1 = -\mu_A^{\mathrm{al}} + \frac{1}{2}\eta_{AA}^{\mathrm{al}} = \left(-\mu_A^{\mathrm{al.el}} + \frac{1}{2}\eta_{AA}^{\mathrm{al}}\right) - \mu_A^{\mathrm{al,num}} = D_A^{1,\mathrm{el}} + D_A^{1,\mathrm{nuc}} \qquad (6\text{-}6)$$

and from Eq.**(4-7)**, the double deprotonation energy is

$$D_{AB}^2 = -\left(\mu_A^{\mathrm{al}} + \mu_B^{\mathrm{al}}\right) + \frac{1}{2}\left(\eta_{AA}^{\mathrm{al}} + \eta_{BB}^{\mathrm{al}}\right) - \eta_{AB}^{\mathrm{al}} = D_A^1 + D_B^1 - \eta_{AB}^{\mathrm{al}}\,. \qquad (6\text{-}7)$$

The results for different deprotonation sites are presented in Fig. VI-5 and Fig. VI-6.

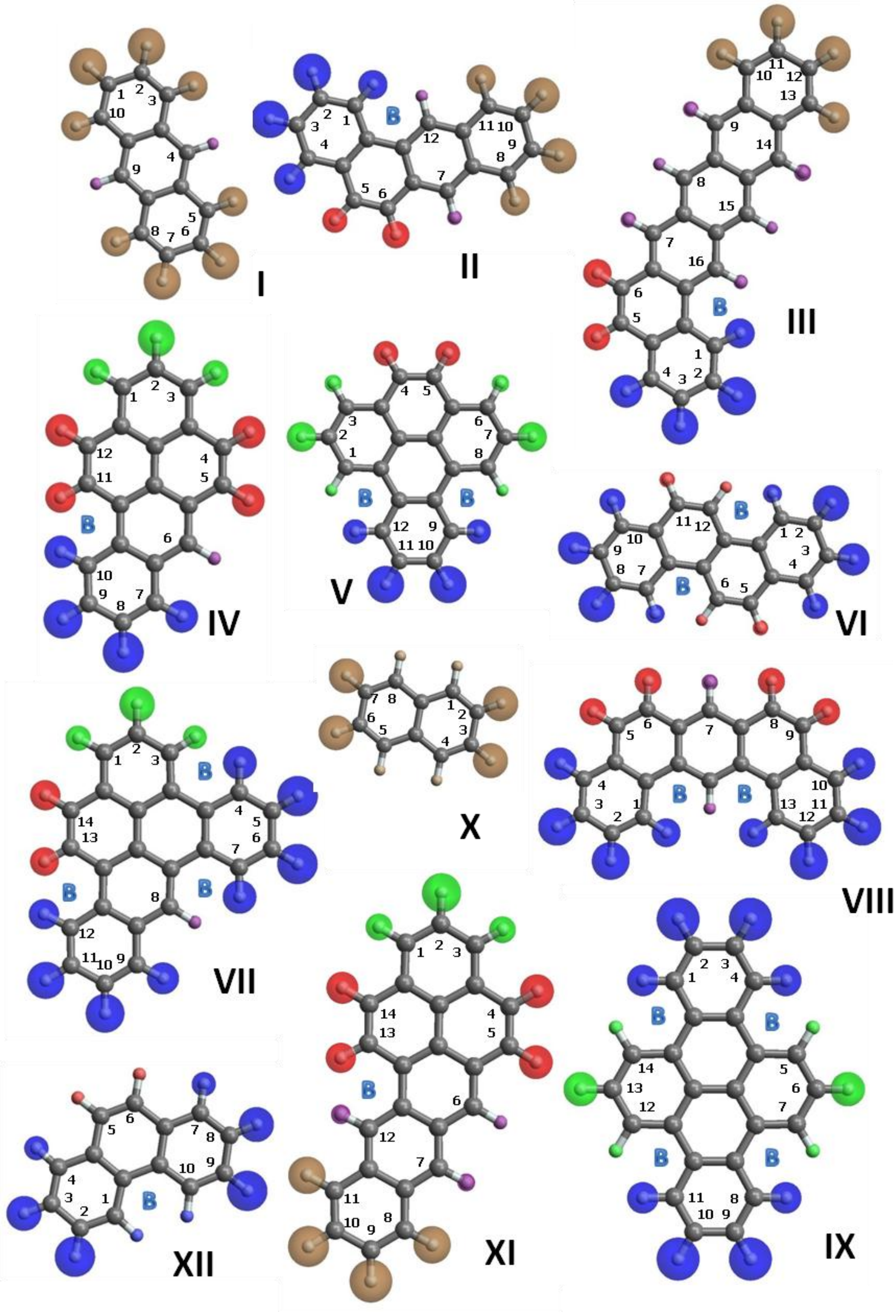

Fig. VI-5. Set of PAHs included in this study. The radius of the sphere at the hydrogen position is proportional to the deprotonation energy at this position. The ratio between the max and min deprotonation energy is 3:1. The numeration of the carbon atom is as IUPAC recommendations.[207] The region color coding is as follows: L -region–purple, K -region–red, N-region–green, M -region–blue, M^0 -region–brown. The bay-region is marked by blue “B”.

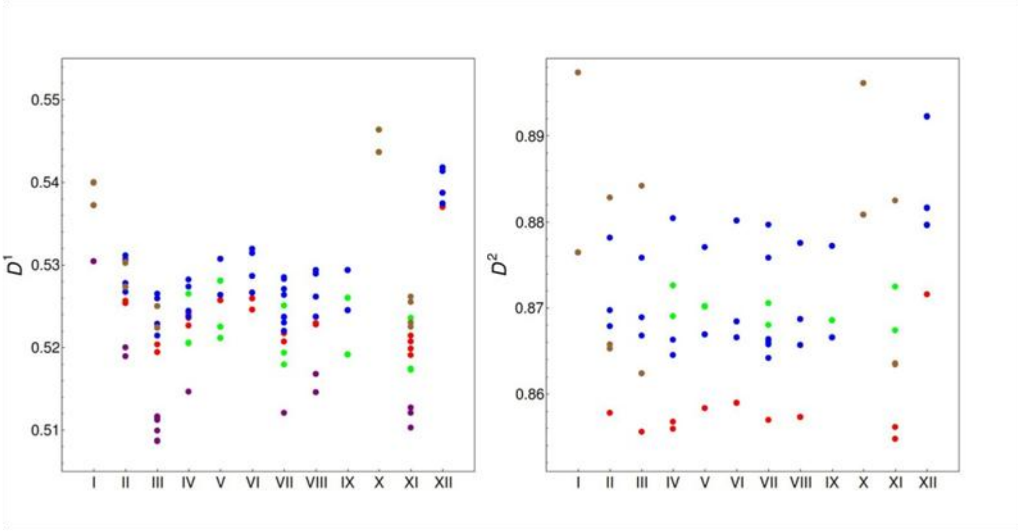

Fig. VI-6. The value of the one and two side deprotonation.

One reason of producing less accurate models in the correlations between the molecular orbital based predictions and carcinogenicity is that these indices used so far are inadequate to characterize the formation of the electrophilic metabolites of PAHs. They are unable to distinguish between the formation of a radical and that of carbocation. Such a distinction is necessary for an improved correlation with experimental facts. In the MCS model[206], three important influences on carcinogenic potency are taken into account: the initial epoxidation of the M -region (M) in competition with reactions on other centers of the PAH molecule, the carbocation formation (C) and the size and solubility (S). In the biomimetic model for enzymatic epoxidation, one of the discussed mechanisms considers the more reactive carbon atom as the beginning side.[206] If we assume that the more reactive carbon should be connected with the preferable deprotonation site, the lowest deprotonation energy may be considered as an indicator of such a place. After inspection of Fig. VI-5 and Fig. VI-6, we can conclude that the lower D_A^1 are for the carbon belonging to the L -region or N-region, eg. carbons at position 7 and 12 in (**II**), or carbon at position 1 in (**V**). In Fig. VI-6 (left panel), we can see that in all molecule the lower deprotonation energy is at the L - or N- region, two type of deactivating regions. The lower D_A^1 in the case of the M -region carbon is related to the initial epoxidation position, e.g. epoxidation of benzo[a]pyrene (**IV**) at its 7,8-position.

The carcinogenicity of (**IV**) is the most studied example of PAHs since many decades ago.[175,208] According to the alchemical deprotonation energy results shown in Fig. VI-5, position 6 of B[a]P is the most preferable site for one-electron oxidation pathway (radical attack). This is in agreement with the experimental evidences.[183] Moreover, the ionization potential for (**IV)** is 7.12 eV which is less than the 7.35 eV (one criteria of radical reaction to occur). From experimental evidences, that epoxidation of (**IV**) occurs at 7, 8 and 9,10 positions.[172] This correlates well with the fact that in the case of double deprotonation, the hydrogens connected with the two bonded carbons are more preferable than with two no-bonded carbons (the former have highest η_{AB}^{al} value than the latter, than probably the lowest deprotonation energy). In Fig. VI-6 (right panel) the lowest value for the blue points (the *M*-region carbons pair) are at 7,8-position and 9,10-position for the (**IV**) . Naphtho[2,3-a]pyrene (**XI**) is also carcinogenic molecule.[206] From the alchemical derivatives point of view, the 6-positions is the most probable site for radical attack, though there is no clear experimental evidence to prove this mechanism.[209] Dibenzo[a,e]pyrene (**VII**) is another PAH caused malignant skin tumor and is a weak to moderate carcinogen.[205,210,211] It has D_{min}^{1} value at position 8, which is the most probable site for radical attack and is in agreement with experiment.[210]

In general the M-region hydrogen with minimal value of D_{A}^{1} can be identified as the hydrogen connected with epoxidation staring carbon in MCS model, when the hydrogen with the lowest value of D_{A}^{1} for whole molecule is connected with the carbon position for the competed reaction in the MCS model.[206]

Another interesting one is the fact that the place of the lowest double deprotonation energy is the K - region if such region is present in the molecule (see red points in Fig. VI-6, right panel).

Based on the excellent relation between the $D^{\mathrm{al,nuc}}$ and $D^{\mathrm{al,el}}$ observed in Sec.IV.E (see Fig. IV-8.b), the electronic component vs. the nuclear component is presented in Fig. VI-7.

It can be noted that for all molecule, the region with maximal values of the nuclear components is observed for the hydrogen in the M - or M^{0} - regions, while minimal values occur for the hydrogens belonging to the L -regions.

It seems that the alchemical derivative can serve as good descriptors of the structural properties of PAHs, but it needs detailed investigation (since they are related with the process which produces anionic systems) of relations between them and the mechanisms of carcinogenic activity at every stage of the metabolite formation.

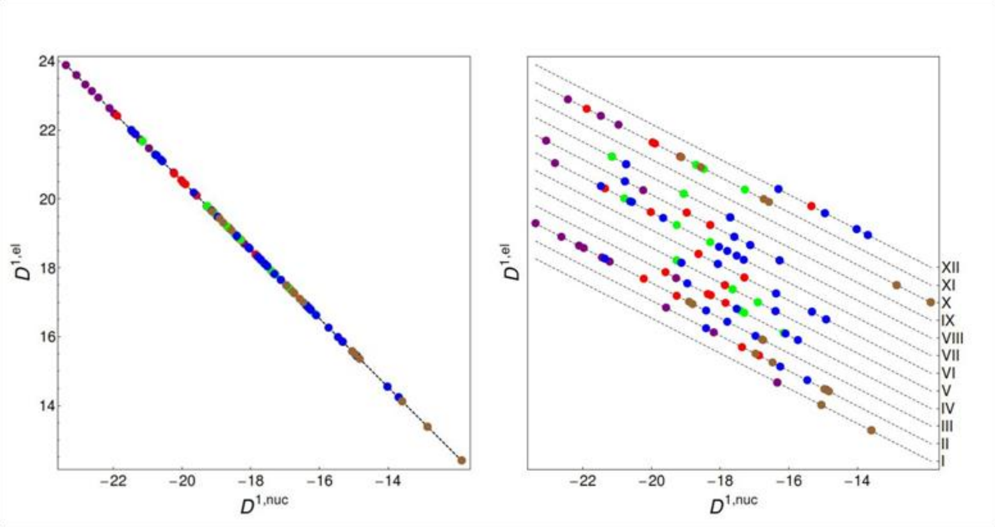

Fig. VI-7. The electronic component vs. the nuclear components of the deprotonation energy (left panel). The values of $D^{al,el}$ for a given molecule are shifted by constants. The molecule numbering is as in Fig. VI-5.

Chapter VII. Conclusion and Summary

Is it possible to predict which molecule is more stable, more reactive? Or based on some knowledge, is it possible to rationalize experimental facts?

Different chemical theories had and have the ambition to answer on these and other questions related to the chemical reactivity. In this thesis, the conceptual density functional theory and the information theory, both use the electronic density (probability of finding the electron in a specific position) are used in the exploring the chemical space. The most fruitful and promising framework so far is probably the conceptual DFT in which one tries to extract from the electronic density relevant concepts and principles that help to understand and predict the chemical behavior of molecules.

The status of the information theory as the tool in the chemical reactivity investigations, stimulated by the fact that density function is ultimately a probability distribution, is still uncertain and in the course of crystallization

One of the main targets undertaken in this work was opening the gate to explore the chemical space through alchemical derivatives. This was done by developing the methodological framework of alchemical derivatives and their chemical reactivity based indices. The crucial conditions for the qualitative and quantitative accuracies of the alchemical predictions were recognized. As illustrative transmutations showing the potential of the method in exploring Chemical Space, some examples of increasing complexity starting with the deprotonation, continuing with the transmutation of the nitrogen molecule, and ending with the substitution of isoelectronic B-N units for C-C units and N units for C-H units in carbocyclic systems were presented. The overall trends observed for the alchemical deprotonation energy prove the usefulness of the alchemical indices as the probe in the chemical reactivity investigations. The results of calculations for the BN derivatives of benzene and pyrene show that this method has great potential for efficient and accurate scanning of Chemical Space. The tentative results for the carcinogenic activity of the PAHs reinforce this opinion.

The information theory based methodological framework necessary for extracting useful information for atomic and molecular systems was deeply examined. The concepts of the transferability and additivity of atoms or functional groups were used as a test in extensive and detail analyses of the electronic density and the shape function as a functional argument for the IT based measures. It was also shown that in the shape representation, the observed trends for the IT measures and complexities are in contradiction with chemical intuition.

It is important to point out that linear correlations are obtained between the kinetic energy and the Fisher information and Onicescu information measures as well as between the atomization energy and the atomization entropy. The analysis of the IT based measures of information planes has shown that the Shannon-Fisher information plane provides “richer” information about the pattern, organization, similarity of molecules than the Shannon-Onicescu and Fisher-Onicescu planes. The final conclusion for this part is that, the IT measures can be used in the chemical reactivity investigation as the source of the information about the pattern, organization, similarity of molecules, while not as direct indicators of their reactivity.

Finally, a support vector machine based models of classification and regression are developed for the carcinogenic effect of polycyclic aromatic hydrocarbons.

An accuracy of 93% of correct classification is achieved using selected structural and molecular descriptors. The correlation coefficient for the predicted versus experimental index of carcinogenicity is about 0.9475.

The used sets of molecule were large enough (in total around 1000 molecules) and diverse to improve the previous understating of subject undertaken in this thesis and to generalize obtained conclusions. The combined results of different methodologies and properties derived from different chemical descriptors provide a guarantee in concluding the behavior/property of molecules while using in many areas of application. Hence, such investigation can serve as a short cut of experimental investigations that are carried out through costly and sophisticated laboratory techniques or even impossible to measure experimentally.

Appendix A. The Compatibility, Accuracy and Deficiency Numbers

To characterize the accuracy qualitatively let us introduce the compatibility number, the accuracy number and the deficiency number for a given series of isomers. Let us explain their meaning through the following example. For a given level of calculations two lists are defined: $\mathbf{p}^{\mathrm{al}}=\left\{p_1^{\mathrm{al}},...,p_{10}^{\mathrm{al}}\right\}, p_i^{\mathrm{al}}=i$, which represents new labels of the isomers numbering the 10 energetically lowest isomers, $D_i^{\mathrm{al}}<D_{i+1}^{\mathrm{al}}$, and $\mathbf{p}^{\mathrm{ver}}=\left\{p_1^{\mathrm{ver}},...,p_{10}^{\mathrm{ver}}\right\}, p_i^{\mathrm{ver}}\in\left\{1,..,10\right\}$, which gives the positions of the 10 lowest vertical transmutation energies for these isomers. If $p_i^{\mathrm{ver}}\neq i$, then the alchemical prediction for the isomers position is inconsistent with the vertical prediction (using the above labels). The *compatibility number* is defined as the number of isomers for which the alchemical position is equal to the vertical position ($p_i^{\mathrm{ver}}=i$). The *accuracy number* is defined as the number of the vertical positions of the lowest isomers which are consistent with the alchemical predictions.

As an example, for all calculations $\mathbf{p}^{\mathrm{al}}$ is simply $\left\{1,2,3,4,5,6,7,8,9,10\right\}$ and for the isomers with three BN moieties $\mathbf{p}^{\mathrm{ver}}=\left\{1,2,4,3,5,6,7,8,9,10\right\}$ (at B3LYP/cc-pCVTZ level). This means that for the isomers with the labels 4 and 3, the alchemical prediction for the position was wrong. The compatibility and accuracy numbers are 8 and 2, respectively. The deficiency number for the given method is defined as the number of the 10 energetically lowest isomers which are not in the reference set, e.g. at HF/STO-3 level, only one isomer is not in the B3LYP/cc-pCVTZ set (see Table IV-8).

Appendix B. Information Theory Measures as the Functionals of the Spin Density and Spin-Density Shape

Relations between the IT measures calculated in the spin-position and the position spaces for the density and shape representations are shortly presented. The relation between these functionals can be obtained with the help of the related functionals of the spin density components, $\rho_\alpha(\mathbf{r})$ and $\rho_\beta(\mathbf{r})$.

Each considered spinor density functional ($S[\rho]$, $I[\rho]$, $D[\rho]$) happens to be a local functional of $\rho(\mathbf{x})$ and $\nabla\rho(\mathbf{x}), F[\rho]=\int f\left(\rho(\mathbf{x}),\nabla\rho(\mathbf{x})\right)d\mathbf{x}$, therefore it is a sum of spin-component functionals

$$F[\rho]=\sum_\kappa\int f\left(\rho(\mathbf{r},\kappa),\nabla\rho(\mathbf{r},\kappa)\right)d\mathbf{r}=F[\rho_\alpha]+F\left[\rho_\beta\right]. \tag{B-1}$$

Analogous relation for the spinless density, Eq.(2-87), can be written formally as

$$F[\rho_\mathrm{N}]=F[\rho_\alpha]+F\left[\rho_\beta\right]+\tilde{F}\left[\rho_\mathrm{N}(\mathbf{r}),\rho_\alpha(\mathbf{r}),\rho_\beta(\mathbf{r})\right], \tag{B-2}$$

where $\tilde{F}\left[\rho_\mathrm{N}(\mathbf{r}),\rho_\alpha(\mathbf{r}),\rho_\beta(\mathbf{r})\right]\equiv F[\rho_\mathrm{N}]-\left(F[\rho_\alpha]+F\left[\rho_\beta\right]\right)$.

The transformation to the shape functionals, Eq.**(5-3)**, goes via the relation

$$F[\zeta]=F\left[N_\zeta\sigma_\zeta\right]. \tag{B-3}$$

In terms of $\sigma_\zeta(\mathbf{q})=\zeta(\mathbf{q})/N_\zeta$, the following relations hold

$$\sigma(\mathbf{r},\kappa)=\frac{1}{N}\rho_\kappa(\mathbf{r})=\frac{N_\kappa}{N}\sigma_\kappa(\mathbf{r}) \text{ , i.e. } \sigma(\mathbf{r},\kappa)\neq\sigma_\kappa(\mathbf{r}) \tag{B-4}$$

$$\sigma_\mathrm{N}(\mathbf{r})=\frac{1}{N}\left(\rho_\alpha(\mathbf{r})+\rho_\beta(\mathbf{r})\right)=\sigma_\alpha(\mathbf{r})\frac{N_\alpha}{N}+\sigma_\beta(\mathbf{r})\frac{N_\beta}{N}=\sum_\kappa\sigma(\mathbf{r},\kappa). \tag{B-5}$$

Basing on these relations, we can pass between different representations and spaces (here N and N_κ are also involved)

$$\begin{tikzcd}
\rho \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & \left\{\rho_\alpha,\rho_\beta\right\} \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & \rho_\mathrm{N} \arrow[d, leftrightarrow] \\
\sigma \arrow[r, leftrightarrow] & \left\{\sigma_\alpha,\sigma_\beta\right\} \arrow[r, leftrightarrow] & \sigma_\mathrm{N}
\end{tikzcd} \tag{B-6}$$

The relations between the Shannon entropies (SEs) in the density representations are

$$S[\rho]=S[\rho_\alpha]+S\left[\rho_\beta\right], \qquad S[\rho_{\mathrm{N}}]=S[\rho]+\tilde{S}_{\mathrm{KL}}[\rho_\alpha,\rho_{\mathrm{N}}]+\tilde{S}_{\mathrm{KL}}\left[\rho_\beta,\rho_{\mathrm{N}}\right], \tag{B-7}$$

where

$$\tilde{S}_{\mathrm{KL}}[\rho_\kappa,\rho_{\mathrm{N}}]\equiv\int\rho_\kappa(\mathbf{r})\ln\frac{\rho_\kappa(\mathbf{r})}{\rho_{\mathrm{N}}(\mathbf{r})}d\mathbf{r}, \tag{B-8}$$

known as the Kullback-Leibler entropy[212].

Due to the properties that $\rho_\kappa(\mathbf{r})/(\rho_\alpha(\mathbf{r})+\rho_\beta(\mathbf{r}))\leq 1$ and $\rho_\kappa(\mathbf{r})\geq 0$, $\tilde{S}_{\mathrm{KL}}[\rho_\kappa,\rho_{\mathrm{N}}]$ is non-positive quantity, so $S[\rho]\geq S[\rho_{\mathrm{N}}]$. Only in the fully spin-polarized case, $\rho_\alpha(\mathbf{r})=\rho_{\mathrm{N}}(\mathbf{r})$, $\rho_\beta(\mathbf{r})=0$, these entropies are equal).

The Fisher informations (FIs) are related as follows

$$I[\rho]=I[\rho_\alpha]+I\left[\rho_\beta\right], \qquad I[\rho_{\mathrm{N}}]=I[\rho]-\int\frac{\rho_\alpha(\mathbf{r})\rho_\beta(\mathbf{r})}{\rho_{\mathrm{N}}(\mathbf{r})}\left|\nabla\ln\left(\frac{\rho_\beta(\mathbf{r})}{\rho_\alpha(\mathbf{r})}\right)\right|^2 d\mathbf{r}. \tag{B-9}$$

so $I[\rho]\geq I[\rho_{\mathrm{N}}]$

The Onicescu informations (OIs) (the disequilibriums) are related as

$$D[\rho]=D[\rho_\alpha]+D\left[\rho_\beta\right], \qquad D[\rho_{\mathrm{N}}]=D[\rho]+2\int\rho_\alpha(\mathbf{r})\rho_\beta(\mathbf{r})d\mathbf{r}, \tag{B-10}$$

so $D[\rho_{\mathrm{N}}]\geq D[\rho]$.

To pass from the density representation to the shape representation, Eq.(B-3) is used,

$$S\left[\sigma_\zeta\right]=S[\zeta]/N_\zeta+\ln N_\zeta, \tag{B-11}$$

$$I\left[\sigma_\zeta\right]=I[\zeta]/N_\zeta, \tag{B-12}$$

$$D\left[\sigma_\zeta\right]=D[\zeta]/N_\zeta^2. \tag{B-13}$$

The relations between the spinor-shape functionals and spinless-shape functionals are more complicated than their counterparts in the density representation. For the SE relations, we have

$$\begin{aligned} S[\sigma]&=\sum_\kappa\frac{N_\kappa}{N}\left(S[\sigma_\kappa]-\ln N_\kappa\right)+\ln N, \\ S[\sigma_{\mathrm{N}}]&=S[\sigma]+\sum_\kappa\frac{N_\kappa}{N}\left(\tilde{S}_{\mathrm{KL}}[\sigma_\kappa,\sigma_{\mathrm{N}}]+\ln\frac{N_\kappa}{N}\right) \\ &=\sum_\kappa\frac{N_\kappa}{N}\left(S[\sigma_\kappa]+\tilde{S}_{\mathrm{KL}}[\sigma_\kappa,\sigma_{\mathrm{N}}]\right), \end{aligned} \tag{B-14}$$

see Eq.(B-8) for $\tilde{S}_{\mathrm{KL}}$. Since $\sigma_\alpha(\mathbf{r})$ is less (greater) than $\sigma_{\mathrm{N}}(\mathbf{r})$ when $\sigma_\alpha(\mathbf{r})$ is less (greater) than $\sigma_\beta(\mathbf{r})$, the sign of $\tilde{S}_{\mathrm{KL}}[\sigma_\kappa,\sigma_{\mathrm{N}}]$ is unpredictable.

The relations for FI are as follows

$$\begin{aligned} I[\sigma] &= \frac{N_\alpha}{N} I[\sigma_\alpha] + \frac{N_\beta}{N} I[\sigma_\beta], \\ I[\sigma_{\mathrm{N}}] &= I[\sigma] - \frac{N_\alpha N_\beta}{N^2} \int \frac{\sigma_\alpha(\mathbf{r})\sigma_\beta(\mathbf{r})}{\sigma_{\mathrm{N}}(\mathbf{r})} \left| \nabla \ln\left(\frac{\sigma_\beta(\mathbf{r})}{\sigma_\alpha(\mathbf{r})}\right) \right|^2 d\mathbf{r}, \end{aligned} \qquad \text{(B-15)}$$

and the relations for OI are

$$\begin{aligned} D[\sigma] &= \frac{1}{N^2}\left(N_\alpha^2 D[\sigma_\alpha] + N_\beta^2 D[\sigma_\beta]\right), \\ D[\sigma_{\mathrm{N}}] &= D[\sigma] + 2\frac{N_\alpha N_\beta}{N^2} \int \sigma_\alpha(\mathbf{r})\sigma_\beta(\mathbf{r})\, d\mathbf{r}. \end{aligned} \qquad \text{(B-16)}$$

Using Eq.(B-7), the relations for the exponential SE in the density representation are

$$H[\rho] = H[\rho_\alpha] H[\rho_\beta], \qquad H[\rho_{\mathrm{N}}] = H[\rho] \exp\left(\sum_\kappa \tilde{S}_{\mathrm{KL}}[\rho_\kappa,\rho_{\mathrm{N}}]\right), \qquad \text{(B-17)}$$

and

$$J[\rho] = (2\pi e) J[\rho_\alpha] J[\rho_\beta], \quad J[\rho_{\mathrm{N}}] = J[\rho] \exp\left(\frac{2}{3}\sum_\kappa \tilde{S}_{\mathrm{KL}}[\rho_\kappa,\rho_{\mathrm{N}}]\right). \qquad \text{(B-18)}$$

In the shape representation, using Eq.(B-14)

$$\begin{aligned} H[\sigma] &= N \prod_\kappa \left(H[\sigma_\kappa]/N_\kappa\right)^{\frac{N_\kappa}{N}}, \\ H[\sigma_{\mathrm{N}}] &= H[\sigma] \prod_\kappa \left(\frac{N_\kappa}{N} \exp\left(\tilde{S}_{\mathrm{KL}}[\sigma_\kappa,\sigma_{\mathrm{N}}]\right)\right)^{\frac{N_\kappa}{N}} \end{aligned} \qquad \text{(B-19)}$$

and

$$\begin{aligned} J[\sigma] &= N\left(\frac{1}{2\pi e}\right)^{1-\frac{N_\alpha N_\alpha}{N^2}} \prod_\kappa \left(N_\kappa^{-2/3} J[\sigma_\alpha]\right)^{\frac{N_\alpha}{N}}, \\ J[\sigma_{\mathrm{N}}] &= J[\sigma] \exp\left(\sum_\kappa \frac{N_\kappa}{N}\left(\tilde{S}_{\mathrm{KL}}[\sigma_\kappa,\sigma_{\mathrm{N}}] + \ln\frac{N_\kappa}{N}\right)\right). \end{aligned} \qquad \text{(B-20)}$$

For the LMC complexity, we have the following relationships

- between $C_{\mathrm{LMC}}[\sigma_{\mathrm{N}}]$ and $C_{\mathrm{LMC}}[\sigma]$, using Eq.(B-16) and Eq.(B-19)

$$C_{\mathrm{LMC}}\left[\sigma_{\mathrm{N}}\right]=C_{\mathrm{LMC}}\left[\sigma\right]\left(\frac{N_{\alpha}N_{\alpha}}{N^{2}}\right)\exp\left(\sum_{\kappa}\tilde{S}_{\mathrm{KL}}\left[\sigma_{\kappa},\sigma_{\mathrm{N}}\right]\right)$$

$$+2H\left[\sigma\right]\left(\frac{N_{\alpha}N_{\alpha}}{N^{2}}\right)\exp\left(\sum_{\kappa}\tilde{S}_{\mathrm{KL}}\left[\sigma_{\kappa},\sigma_{\mathrm{N}}\right]\right)\left(\frac{N_{\alpha}N_{\beta}}{N^{2}}\int\sigma_{\alpha}\left(\mathbf{r}\right)\sigma_{\beta}\left(\mathbf{r}\right)d\mathbf{r}\right), \quad \text{(B-21)}$$

- between $C_{\mathrm{LMC}}\left[\sigma_{\mathrm{N}}\right]$ and $C_{\mathrm{LMC}}\left[\rho_{\mathrm{N}}\right]$, using Eqs. (B-11)

$$C_{\mathrm{LMC}}\left[\sigma_{\mathrm{N}}\right]=D\left[\sigma_{\mathrm{N}}\right]H\left[\sigma_{\mathrm{N}}\right]=\frac{1}{N}D\left[\rho_{\mathrm{N}}\right]\exp\left(S\left[\rho_{\mathrm{N}}\right]\right)^{\frac{1}{N}}$$

$$=\frac{H\left[\rho_{N}\right]^{\frac{1-N}{N}}}{N}C_{\mathrm{LMC}}\left[\rho_{\mathrm{N}}\right], \quad \text{(B-22)}$$

- between $C_{\mathrm{LMC}}\left[\rho_{\mathrm{N}}\right]$ and $C_{\mathrm{LMC}}\left[\rho\right]$, using Eq. (B-10) and Eq.(B-17)

$$C_{\mathrm{LMC}}\left[\rho_{\mathrm{N}}\right]=\left(C_{\mathrm{LMC}}\left[\rho\right]+2\int\rho_{\alpha}\left(\mathbf{r}\right)\rho_{\beta}\left(\mathbf{r}\right)d\mathbf{r}\right)\exp\left(\sum_{\kappa}\tilde{S}_{\mathrm{KL}}\left[\rho_{\alpha},\rho_{\mathrm{N}}\right]\right). \quad \text{(B-23)}$$

The realtions for the FS complexity can be obtained in a similar way.

It should be noted that for the spin-compensated system, where $\sigma_{\alpha}\left(\mathbf{r}\right)=\sigma_{\beta}\left(\mathbf{r}\right)=\sigma_{\mathrm{N}}\left(\mathbf{r}\right)$, and $\rho_{\alpha}\left(\mathbf{r}\right)=\rho_{\beta}\left(\mathbf{r}\right)=\rho_{\mathrm{N}}\left(\mathbf{r}\right)/2$ all these relations are simpler due to the fact that

$$\tilde{S}_{\mathrm{KL}}\left[\sigma_{N},\sigma_{\mathrm{N}}\right]=0 \quad \text{(B-24)}$$

and

$$\tilde{S}_{\mathrm{KL}}\left[\rho_{\kappa},\rho_{\mathrm{N}}\right]=\tilde{S}_{\mathrm{KL}}\left[\rho_{\mathrm{N}}/2,\rho_{\mathrm{N}}\right]=\frac{1}{2}\int\rho_{\mathrm{N}}\left(\mathbf{r}\right)\ln\frac{\rho_{\mathrm{N}}}{2\rho_{\mathrm{N}}}d\mathbf{r}=-\frac{N\ln 2}{2}. \quad \text{(B-25)}$$

A compilation of all relations presented in this Appendix for spin-compensated systems is displayed in Table V-1.

Appendix C. Support Vector Machines

Support Vector Machines (SVMs)[213-215] are a new generation of supervised machine learning algorithms (labeled data based algorithms) that are aimed for classification, regression, and novelty detection. They are an extension of the nonlinear models of the generalized portrait algorithm developed by Vladimir Vapnik [213]. SVMs are discriminative classifier algorithms to produce an optimal hyperplane which categorizes new examples. It is based on the maximum margin principle to achieve the maximum separation between two classes or maximizing the distance from hyperplane to the nearest data point. For *n* labeled training data set $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ with $\mathbf{x} \in \Re^m$ as *m*-dimensional input vector and $y_i \in \{-1, +1\} \equiv \{\mathrm{F}, \mathrm{T}\}$ two-class variable for each, the separating hyperplane is defined by

$$y_i \left(\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x}_i + b \right) \geq 1 \tag{C-1}$$

which is equivalent to the constraints

$$\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x}_i + b \geq 1 \;\text{ for }\; y_{\mathrm{i}} = +1 \tag{C-2}$$

$$\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x}_i + b \leq -1 \;\text{ for }\; y_{\mathrm{i}} = -1 . \tag{C-3}$$

where **w** is the weighted vector perpendicular to the optimal separation hyperplane (the decision boundary) for all attributes. The real number *b* represents the bias (intercept). This optimal separation hyperplane is the set of points that satisfy $\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b = 0$. The other two parallel hyperplanes that define the margin are $\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b = -1$ and $\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b = 1$ and no training points fall between them. Note that a training set is a set of examples used for learning with known target value $\{-1, +1\}$. Support vectors are training points lie on these two hyperplanes. Decision boundary is determined only by those support vectors. The distance between these two hyperplanes (P1 and P2 in Fig. C-1(a)) is called the margin $2 / \|\mathbf{w}\|$. Hence, in order to maximize the margin, one has to minimize $\|\mathbf{w}\|$:

$$\min_{w,b} \left[\frac{1}{2} \|\mathbf{w}\|^2 \right] , \tag{C-4}$$

with respect to the constraint Eq.(C-1). The minimization in Eq. (C-4) is equivalent[216] to maximizing the Wolfe dual function in terms of Lagrange multipliers α_{i}

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(\mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{x}_j \right), \tag{C-5}$$

with respect to the variables αi subject to constraint $0 \le \alpha_i \le C$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$.

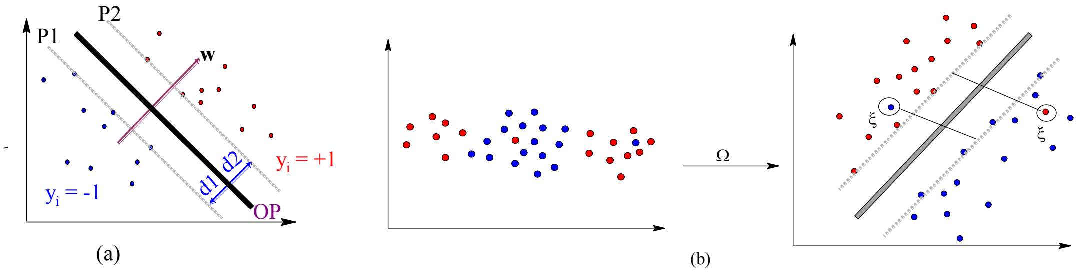

Fig. C-1. (a) Two linearly separable classes in which P1 and P2 are hyperplanes while d1 and d2 are the shortest distance to the closest negative (y_i = -1) and positive (y_i = +1) class respectively. The sum of distances d1 and d2 is called the margin. OP is the optimal hyperplane. Points on the P1 and P2 are support vectors. The normal vector perpendicular to the optimal hyperplane is the weighting factor vector, **w**. (b) Linearly inseparable (nonlinear) data points are mapped into linearly separable higher *n*-dimensional feature space using kernel function, Ω. The encircled data points are outliners labeled as slack variables ξ.

SVM classifier is based on the concept of decision planes that define decision boundaries by mapping linearly inseparable input data into a higher dimensional Hilbert space (usually known as feature space) such that the classification problem becomes simpler or linearly separable in the feature space. As shown in Fig. C-1(b) there is no hyperplane to classify the points into two pure classes (red and blue points) exactly. In such cases, to handle outliners, it is important to allow few points to be outside the margin, such treatment is called the soft margin[217]. Introducing a non-negative number, the slack variable ξ, Eq.(C-4) is modified to

$$\min_{w,b,\xi_i} \left[\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \right], \tag{C-6}$$

subject to the constraint

$$y_i \left(\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x}_i + b \right) \ge 1 - \xi_i \quad \forall i,\ \xi_i \ge 0. \tag{C-7}$$

Here C > 0 is the cost (penalizing) parameter. It defines the maximizing margin and minimizes the amount of ξ. The slack variable allows an example to be in the margin, i.e. $0 \le \xi_i \le 1$ whereas ξ_i > 1 represent misclassification. Choosing low value of C results in reducing the risk of overfitting on the training sample. The value of C is determined during the optimization procedure.

Using Lagrange multiplier, minimize Eq.(C-6) with the constraint $0 \le \alpha_i \le C$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$ [218] which leads to the weight factor

$$\mathbf{w}^* = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i \,. \tag{C-8}$$

For a linear classifier this yields a decision function of the form

$$f(\mathbf{x}) = \operatorname{sign}\left(\mathbf{w}^{*\mathrm{T}} \cdot \mathbf{x} + b\right) = \operatorname{sign}\left(\sum_{i}^{n} y_i \alpha_i \mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{x} + b\right), \tag{C-9}$$

while for nonlinear case

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{m} y_i \alpha_i K\left(\mathbf{x}_i, \mathbf{x}\right) + b\right). \tag{C-10}$$

In the sum only some α_i are nonzero. The function $K(\mathbf{x}_i, \mathbf{x})$ is a kernel function (satisfying the Mercer theorem[219] that represent a dot product of input data points mapped into the higher dimensional feature space by Ω

$$K\left(\mathbf{x}_i, \mathbf{x}\right) = \Omega\left(\mathbf{x}_i\right)^{\mathrm{T}} \cdot \Omega\left(\mathbf{x}\right). \tag{C-11}$$

Here a nonlinear input space is transformed by Ω (no need to know) into high dimensional featured space (see Fig. C-1(b)). The following kernel functions[213,220] in Eq.(C-12) -(C-15) represent the linear, polynomial, Gaussian radial basis function (RBF) and Laplace function respectively

$$K\left(\mathbf{x}_i, \mathbf{x}\right) = \left\langle \mathbf{x}_i, \mathbf{x} \right\rangle = \mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{x}, \tag{C-12}$$

$$K\left(\mathbf{x}_i, \mathbf{x}\right) = \left(\gamma \left\langle \mathbf{x}_i, \mathbf{x} \right\rangle + 1\right)^d, \tag{C-13}$$

$$K\left(\mathbf{x}_i, \mathbf{x}\right) = \exp\left(-\gamma \left\| \mathbf{x}_i - \mathbf{x} \right\|^2\right), \tag{C-14}$$

$$K\left(\mathbf{x}_i, \mathbf{x}\right) = \exp\left(-\gamma \left\| \mathbf{x}_i - \mathbf{x} \right\|\right). \tag{C-15}$$

Here γ is a parameter and *d* is the degree of the polynomial. The radial kernel is the most popular choice in SVM because of its localized and finite response across the entire range of the real x-axis.

SVMs can also be applied for regression. The aim of SVM regression is to find a function $f(\mathbf{x})$ that has the least deviation between predicted and actual (experimentally observed) responses for all training examples. For simple linearly dependent data, the regression function is

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{C-16}$$

SVM regression minimizes the combination of the training error and a regularization term (controls the complexity of featured space), collectively called the generalization error bound, to achieve higher generalization performance. The training error is computed by the ε-insensitive loss function,[213] which describes how the estimated function deviated from the true one,

$$L_\varepsilon\left(\mathrm{y}, f(\mathbf{x})\right) := \left|\mathrm{y} - f(\mathbf{x})\right|_\varepsilon = \min\left(0, \left|\mathrm{y} - f(\mathbf{x})\right| - \varepsilon\right), \tag{C-17}$$

and generates a model representing a tube with radius ε fitted to the data. This loss function ignores errors that are smaller than a certain threshold ε (it only penalizes errors between the estimated values and actual values greater than ε).The regularization term is done by minimizing the norm of the weighted factor **w**. For *n* training set of examples $\{\mathbf{x}_i, \mathrm{y}_i\}_{i=1}^{n}$ with input vector $\mathbf{x}_i \in \Re^n$ and output vector y$_i$, the convex optimization problem is

$$\min_{\mathbf{w}, b, \xi_i, \xi_i^*} \left[\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^* \right) \right], \tag{C-18}$$

subject to the constraints

$$\begin{aligned} &\mathrm{y}_i - (\mathbf{w} \cdot \mathbf{x} + b) \leq \varepsilon + \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, 2, \ldots, n \\ &(\mathbf{w} \cdot \mathbf{x} + b) - \mathrm{y}_i \leq \varepsilon + \xi_i^*, \ \ \xi_i^* \geq 0 \end{aligned} \tag{C-19}$$

The constant *C* controls the tradeoff between the flatness of the regress function. The two slack variables ξ and ξ* are needed to specify the upper and the lower training errors subject to an error tolerance ε. The objective function has to be minimized until almost all data points (if ξ or ξ* exists) located inside the tube. Hence, the estimate regression function will be

$$f(\mathbf{x}) = \sum_{i=1}^{n} \left(\xi_i^* - \xi_i \right) K(\mathbf{x}_i, \mathbf{x}) + b, \tag{C-20}$$

with kernel $K(\mathbf{x}_i, \mathbf{x})$ defined in Eq.(C-11) while *b* is the bias. As in the classification case, a dual form using the Lagrange multiplier can be constructed.

Content of Supporting Information Disc (SID)

Chapter IV	SID_Chapter_IV
Chapter V	SID_Chapter_V
Chapter VI	SID_Chapter_VI_01 and SID_Chapter_VI_02

List of Publications

R. Balawender, M. A. Welearegay, M. Lesiuk, F. De Proft, and P. Geerlings,
"Exploring Chemical space with the Alchemical Derivatives"
J. Chem. Theory Comput., 2013,*9*(12), 5327–5340.

M. A. Welearegay, R. Balawender and A. Holas
"Information and complexity measures in the molecular reactivity studies"
Physical Chemistry Chemical Physics "Accepted" for publication.

M. A. Welearegay, R. Balawender and A. Holas
"Substituent effects on the information theory based indices. Group properties"
in preparation.

M. A. Welearegay, R. Balawender and P.W. Ayers
"SVM model for predicting carcinogenicity of polycyclic aromatic hydrocarbons and derivatives"
in preparation.

List of Oral and Poster presentations

1. A poster presentation entittled "*Information theory based indices in the Spin Density Functional Theory*" at the 14th International Density Functional Theory Conference (DFT 2011), Athens, Greece, August 29-September 2, 2011.
2. A poster presentation entitled "*Classifying and quantifying carcinogenicity of polycyclic aromatic hydrocarbons: support vector machines*" at the Current Trends in Theoretical Chemistry VI" conference, Krakow, Poland; 01-05 September, 2013.
3. Oral presentation on "*Information theory as a tool in predicting carcinogenic properties of molecules*" on Reporting and Training Session Users ICM , March (23-26), 2011, Bedlewo , Poland
4. Four seminar oral presentations (from 2010-2013) at the Institute of Physical Chemistry, ICHF PAS, Warsaw, Poland

List of Research Stays

McMaster university under supervision of Prof. Paul W. Ayers,

Hamilton, Canada (March 2012 - September 2012).

I was enrolled in developing new approaches for predicting the potency of carcinogens

Vrije Universiteit Brussel under supervision of Prof. Paul Geerlings,

Brussels, Belgium. (December 2012 - June 2013).

I was working on determination of molecular reactivity using the conceptual density functional theory.

References

[1] I. V. Tetko, Drug Discov Today **10**, 1497 (2005).

[2] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk and V. V. Prokopenko, J Comput Aided Mol Des **19**, 453 (2005).

[3] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. (Courier Dover Publications, 1996).

[4] F. Jensen, *Introduction to computational chemistry*. (John Wiley & Sons, Chichester, England; Hoboken, 2007).

[5] T. Veszprémi and M. Fehér, *Quantum chemistry: fundamentals to applications*. (Kluwer Academic/Plenum, Dordrecht; New York, 1999).

[6] L. Piela, *Ideas of quantum chemistry*. (Elsevier, Amsterdam; Boston DA, 2007).

[7] P. Atkins and J. d. Paula, *Atkins' Physical Chemistry*. (Oxford University Press, 2010).

[8] J. C. Slater, Phys Rev **36**, 57 (1930).

[9] S. F. Boys, Proc Roy Soc A **200**, 542 (1950).

[10] C. J. Cramer, *Essentials of computational chemistry*. (Wiley, Chichester, 2004).

[11] K. A. Peterson and T. H. D. Jr, J Chem Phys **117**, 10548 (2002).

[12] T. Helgaker, J. Olsen and P. Jorgensen, *Molecular Electronic-Structure Theory*. (John Wiley & Sons, 2013).

[13] K. I. Ramachandran, G. Deepa and K. Namboori, *Computational Chemistry and Molecular Modeling: Principles and Applications*. (Springer, 2008).

[14] E. G. Lewars, *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*, 2nd ed. (Springer, Dordrecht Netherlands ; London ; New York, 2011).

[15] D. Young, *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*. (John Wiley & Sons, 2004).

[16] S. R. Langhoff and E. R. Davidson, Int J Quantum Chem **8**, 61 (1974).

[17] D. Xiao, W. Yang and D. N. Beratan, J Chem Phys **129**, 044106 (2008).

[18] B. C. Rinderspacher, J. Andzelm, A. Rawlett, J. Dougherty, D. N. Beratan and W. Yang, J Chem Theory Comput **5**, 3321 (2009).

[19] A. Franceschetti and A. Zunger, Nature **402**, 60 (1999).

[20] C. M. Dobson, Nature **432**, 824 (2004).

[21] A. J. Cohen, P. Mori-Sanchez and W. Yang, Chem Rev **112**, 289 (2012).

[22] A. Savin and H. J. Flad, Int J Quantum Chem **56**, 327 (1995).

[23] P. M. W. Gill and R. D. Adamson, Chem Phys Lett **261**, 105 (1996).

[24] P. M. W. Gill, R. D. Adamson and J. A. Pople, Mol Phys **88**, 1005 (1996).

[25] T. Leininger, H. Stoll, H. J. Werner and A. Savin, Chem Phys Lett **275**, 151 (1997).

[26] Y. Tawada, T. Tsuneda, S. Yanagisawa, T. Yanai and K. Hirao, J Chem Phys **120**, 8425 (2004).

[27] J. Toulouse, F. Colonna and A. Savin, Phys Rev A **70** (2004).

[28] A. V. Krukau, G. E. Scuseria, J. P. Perdew and A. Savin, J Chem Phys **129**, 124103 (2008).

[29] E. Rebolini, A. Savin and J. Toulouse, Mol Phys **111**, 1219 (2013).

[30] R. Kar, J. W. Song and K. Hirao, J Comput Chem **34**, 958 (2013).

[31] J. P. Dahl and J. Avery, *Local Density Approximations in Quantum Chemistry and Solid State Physics*, 1984 edition ed. (Springer, New York, 1984).

[32] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*. (Oxford University Press, 1989).

[33]M. Gell-Mann and K. A. Brueckner, Phys Rev **106**, 364 (1957).
[34]S. H. Vosko and L. Wilk, Phys Rev B **22**, 3812 (1980).
[35]O. A. von Lilienfeld and M. E. Tuckerman, J Chem Phys **125**, 154104 (2006).
[36]O. A. von Lilienfeld and M. E. Tuckerman, J Chem Theory Comput **3**, 1083 (2007).
[37]D. Sheppard, G. Henkelman and O. A. von Lilienfeld, J Chem Phys **133**, 084104 (2010).
[38]P. W. Ayers, J. S. M. Anderson and L. J. Bartolotti, Int J Quantum Chem **101**, 520 (2005).
[39]R. G. Parr, Ann Rev Phys Chem **34**, 631 (1983).
[40]P. Geerlings, F. De Proft and W. Langenaeker, Chem Rev **103**, 1793 (2003).
[41]A. J. Cohen, P. Mori-SÃ¡nchez and W. Yang, J Chem Phys **126** (2007).
[42]R. Balawender and P. Geerlings, J Chem Phys **123**, 124102 (2005).
[43]E. S. Kryachko and E. V. Ludeña, *Energy density functional theory of many-electron systems*. (Kluwer Academic, Dordrecht; Boston, 1990).
[44]E. R. D.-. Davidson, *Reduced Density Matrices in Quantum Chemistry*. (Academic Press, 1976).
[45]M. Wang, X. Hu, D. N. Beratan and W. Yang, J Am Chem Soc **128**, 3228 (2006).
[46]S. Keinan, X. Q. Hu, D. N. Beratan and W. T. Yang, J Phys Chem A **111**, 176 (2007).
[47]M. d'Avezac and A. Zunger, J Phys Condens Matter **19**, 402201 (2007).
[48]P. Kirkpatrick and C. Ellis, Nature **432**, 823 (2004).
[49]C. Lipinski and A. Hopkins, Nature **432**, 855 (2004).
[50]P. Ertl, J Chem Inf Comp Sci **43**, 374 (2003).
[51]G. H. Johannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen and J. K. Norskov, Phys Rev Lett **88**, 255506 (2002).
[52]O. A. von Lilienfeld, R. D. Lins and U. Rothlisberger, Phys Rev Lett **95**, 153002 (2005).
[53]H. Chermette, J Comput Chem **20**, 129 (1999).
[54]J. L. Gazquez, J Mex Chem Soc **52**, 3 (2008).
[55]P. Geerlings and F. De Proft, Phys Chem Chem Phys **10**, 3028 (2008).
[56]A. Cedillo, R. Contreras, M. Galvan, A. Aizman, J. Andres and V. S. Safont, J. J Phys Chem A **111**, 2442 (2007).
[57]S. Liu, T. Li and P. W. Ayers, J Chem Phys **131**, 114106 (2009).
[58]P. W. Ayers, S. B. Liu and T. L. Li, Chem Phys Lett **480**, 318 (2009).
[59]Z. Boisdenghien, C. Van Alsenoy, F. De Proft and P. Geerlings, J Chem Theory Comput **9**, 1007 (2013).
[60]W. Yang, A. J. Cohen, F. De Proft and P. Geerlings, J Chem Phys **136**, 144110 (2012).
[61]N. Sablon, F. De Proft, M. Sola and P. Geerlings, Phys Chem Chem Phys **14**, 3960 (2012).
[62]N. Sablon, F. De Proft and P. Geerlings, J Phys Chem Lett **1**, 1228 (2010).
[63]N. Sablon, F. De Proft and P. Geerlings, Chem Phys Lett **498**, 192 (2010).
[64]N. Sablon, F. De Proft, P. W. Ayers and P. Geerlings, J Chem Theory Comput **6**, 3671 (2010).
[65]P. W. Ayers, Faraday Discuss **135**, 161 (2007).
[66]P. W. Ayers and R. G. Parr, J Am Chem Soc **123**, 2007 (2001).
[67]N. D. Mermin, Phys Rev **137**, A1441 (1965).
[68]R. Balawender, arXiv:1212.1367 ePrint archive (http://arxiv.org/abs/1212.1367), (2012).
[69]J. P. Perdew, R. G. Parr, M. Levy and J. L. Balduz, Phys Rev Lett **49**, 1691 (1982).
[70]M. H. Cohen, M. V. Ganduglia-Pirovano and J. Kudrnovský, J Chem Phys **103**, 3543 (1995).
[71]R. Balawender and P. Geerlings, J Chem Phys **114**, 682 (2001).
[72]R. Balawender, F. D. Proft and P. Geerlings, J Chem Phys **114**, 4441 (2001).

[73] L. Komorowski and P. Ordon, Theor Chem Acc **105**, 338 (2001).

[74] R. G. Parr and R. G. Pearson, J Am Chem Soc **105**, 7512 (1983).

[75] W. Yang and R. G. Parr, Proc Natl Acad Sci USA **82**, 6723 (1985).

[76] R. G. Pearson, J Chem Edu **45**, 581 (1968).

[77] R. G. Pearson, Acc Chem Res **26**, 250 (1993).

[78] K. Fukui, Science **218**, 747 (1982).

[79] R. R. Contreras, P. Fuentealba, M. Galvan and P. Perez, Chem Phys Lett **304**, 405 (1999).

[80] R. G. Parr, L. Von Szentpaly and S. B. Liu, J Am Chem Soc **121**, 1922 (1999).

[81] M. Galvan, A. Vela and J. L. Gazquez, J Phys Chem **92**, 6470 (1988).

[82] M. Galvan and R. Vargas, J Phys Chem **96**, 1625 (1992).

[83] P. Perez, E. Chamorro and P. W. Ayers, J Chem Phys **128**, 204108 (2008).

[84] E. Chamorro, P. Perez, M. Duque, F. De Proft and P. Geerlings, J Chem Phys **129**, 064117 (2008).

[85] R. F. Nalewajski and R. G. Parr, J Chem Phys **77**, 399 (1982).

[86] N. K. Ray and R. G. Parr, J Chem Phys **73**, 1334 (1980).

[87] M. Lesiuk, R. Balawender and J. Zachara, J Chem Phys **136**, 034104 (2012).

[88] C. H. Depuy, S. Gronert, S. E. Barlow, V. M. Bierbaum and R. Damrauer, J Am Chem Soc **111**, 1968 (1989).

[89] A. D. Becke, Phys Rev A **38**, 3098 (1988).

[90] C. T. Lee, W. T. Yang and R. G. Parr, Phys Rev B **37**, 785 (1988).

[91] J. P. Perdew, K. Burke and M. Ernzerhof, Phys Rev Lett **77**, 3865 (1996).

[92] T. H. Dunning, J Chem Phys **90**, 1007 (1989).

[93] D. E. Woon and T. H. Dunning, J Chem Phys **98**, 1358 (1993).

[94] D. E. Woon and T. H. Dunning, J Chem Phys **100**, 2975 (1994).

[95] F. de Proft, N. Sablon, D. J. Tozer and P. Geerlings, Faraday Discuss **135**, 151 (2007).

[96] Y. Zhang, Z. H. Li and D. G. Truhlar, J Chem Theory Comput **3**, 593 (2007).

[97] I. Langmuir, J Am Chem Soc **41**, 1543 (1919).

[98] R. Thissen, O. Witasse, O. Dutuit, C. S. Wedlund, G. Gronoff and J. Lilensten, Phys Chem Chem Phys **13**, 18264 (2011).

[99] W. Weltner and R. J. Vanzee, Chem Rev **89**, 1713 (1989).

[100] A. Palma, L. Sandoval, K. Churyumov, V. Chavushyan and A. Berezhnoy, Int J Quantum Chem **107**, 2650 (2007).

[101] M. W. Wong, R. H. Nobes, W. J. Bouma and L. Radom, J Chem Phys **91**, 2971 (1989).

[102] M. N. Hughes, Bba-Bioenergetics **1411**, 263 (1999).

[103] F. Jensen, Wires Comput Mol Sci **3**, 273 (2013).

[104] M. J. D. Bosdet and W. E. Piers, Can J Chem **87**, 8 (2009).

[105] P. G. Campbell, A. J. Marwitz and S. Y. Liu, Angew Chem **51**, 6074 (2012).

[106] R. Islas, E. Chamorro, J. Robles, T. Heine, J. C. Santos and G. Merino, Struct Chem **18**, 833 (2007).

[107] J. S. Lu, S. B. Ko, N. R. Walters, Y. Kang, F. Sauriol and S. Wang, Angew Chem **52**, 4544 (2013).

[108] X. Y. Wang, H. R. Lin, T. Lei, D. C. Yang, F. D. Zhuang, J. Y. Wang, S. C. Yuan and J. Pei, Angew Chem **52**, 3117 (2013).

[109] M. J. S. Dewar and R. Dietz, J Chem Soc, 2728 (1959).

[110] M. J. S. Dewar, J. Hashmall and V. P. Kubba, J Org Chem **29**, 1755 (1964).

[111] M. J. Bosdet, W. E. Piers, T. S. Sorensen and M. Parvez, Angew Chem **46**, 4940 (2007).

[112] T. Kar, J. Pattanayak and S. Scheiner, J Phys Chem A **107**, 8630 (2003)

[113]S. Sirimulla, *Drug Design, Molecular Modelling, and QSAR Studies of Antimalarial Mefloquine and Artemisinin Derivatives*. (ProQuest, 2007).

[114]C. Acharya, P. R. Seo, J. E. Polli and A. D. Mackerell, Jr., Mol Pharm **5**, 818 (2008).

[115]P. Darpan, *Competition Science Vision*. (Pratiyogita Darpan, 1999).

[116]R. Chang, *Physical Chemistry for the Biosciences*. (University Science Books, 2005).

[117]F. Deproft, W. Langenaeker and P. Geerlings, J Phys Chem **97**, 1826 (1993).

[118]L. Pauling, *The Nature of the Chemical Bond: An Introduction to Modern Structural Chemistry*. (Cornell University Press, Ithaca, NY, 1960).

[119]M. R.S., J Chem Phys **2**, 782 (1934).

[120]S. B. Sears, R. G. Parr and U. Dinur, Isr J Chem **19**, 165 (1980).

[121]R. F. Nalewajski, J Math Chem **43**, 780 (2008).

[122]C. P. Panos, K. C. Chatzisavvas, C. C. Moustakidis, N. Nikolaidis, S. E. Massen and K. D. Sen, in *Statistical Complexity*, edited by K. D. Sen (Springer Netherlands, 2011), p. 49.

[123]G. Maroulis, M. Sana and G. Leroy, Int J Quantum Chem **19**, 43 (1981).

[124]A. M. Simas, A. J. Thakkar and V. H. Smith, Int J Quantum Chem **24**, 527 (1983).

[125]S. R. Gadre, S. B. Sears, S. J. Chakravorty and R. D. Bendale, Phys Rev A **32**, 2602 (1985).

[126]M. H. Ho, R. P. Sagar, V. H. Smith and R. O. Esquivel, J Phys B-At Mol Opt Phys **27**, 5149 (1994).

[127]M. H. Ho, R. P. Sagar, J. M. Perezjorda, V. H. Smith and R. O. Esquivel, Chem Phys Lett **219**, 15 (1994).

[128]M. Ho, V. H. Smith, D. F. Weaver, C. Gatti, R. P. Sagar and R. O. Esquivel, J Chem Phys **108**, 5469 (1998).

[129]P. Geerlings and A. Borgoo, Phys Chem Chem Phys **13**, 911 (2011).

[130]R. F. Nalewajski, J Math Chem **45**, 607 (2009).

[131]L. Tarko, J Math Chem **49**, 2330 (2011).

[132]R. F. Nalewajski, Mol Phys **104**, 365 (2006).

[133]R. F. Nalewajski, Mol Phys **104**, 3339 (2006).

[134]R. F. Nalewajski, Int J Quantum Chem **109**, 2495 (2009).

[135]R. F. Nalewajski, J Math Chem **45**, 709 (2009).

[136]R. F. Nalewajski, J Math Chem **49**, 546 (2011).

[137]M. Molina-Espiritu, R. O. Esquivel, J. C. Angulo and J. S. Dehesa, Entropy **15**, 4084 (2013).

[138]M. Molina-Espiritu, R. O. Esquivel, J. C. Angulo, J. Antolin and J. S. Dehesa, J Math Chem **50**, 1882 (2012).

[139]M. Molina-Espiritu, R. O. Esquivel, J. C. Angulo, J. Antolin, C. Iuga and J. S. Dehesa, Int J Quantum Chem **113**, 2589 (2013).

[140]A. Borgoo, P. Jaque, A. Toro-Labbe, C. V. Alsenoy and P. Geerlings, Phys Chem Chem Phys **11**, 476 (2009).

[141]J. C. Angulo, J. Antolin and R. O. Esquivel, in *Statistical Complexity*, edited by K. D. Sen (Springer Netherlands, 2011), p. 167.

[142]R. O. Esquivel, J. C. Angulo, J. Antolin, J. S. Dehesa, S. Lopez-Rosa and N. Flores-Gallegos, Phys Chem Chem Phys **12**, 7108 (2010).

[143]K. D. Sen, *Statistical Complexity: Applications in Electronic Structure*. (Springer, 2011).

[144]O. Onicescu, **263**, 841 (1966).

[145]I. Bialynickibirula and J. Mycielski, Commun Math Phys **44**, 129 (1975).

[146]W. Kutzelnigg, Angew Chem **12**, 546 (1973).

[147]P. K. Acharya, L. J. Bartolotti, S. B. Sears and R. G. Parr, Proc Natl Acad Sci USA **77**, 6978 (1980).

[148] S. Liu, J Chem Phys **126**, 244103 (2007).
[149] P. K. Bhatia, Inf Sci **97**, 233 (1997).
[150] S. M. Taheri and R. Azizi, Inf Sci **177**, 3871 (2007).
[151] M. Alipour and A. Mohajeri, Mol Phys **110**, 403 (2012).
[152] K. C. Chatzisavvas and C. P. Panos, Int J Mod Phys E **14**, 653 (2005).
[153] R. López-Ruiz, H. L. Mancini and X. Calbet, Phys Lett A **209**, 321 (1995).
[154] J. C. Angulo, J. AntolÃn and K. D. Sen, Phys Lett A **372**, 670 (2008).
[155] K. D. Sen, J. Antolin and J. C. Angulo, Phys Rev A **76**, 032502 (2007).
[156] N. H. March, Phys Lett A **113**, 476 (1986).
[157] A. Holas and N. H. March, Phys Rev A **44**, 5521 (1991).
[158] V. A. Savin, A. D. Becke, J. Flad, R. Nesper, H. Preuss and H. G. Von Schnering, Angew Chem **103**, 421 (1991).
[159] H. Eschrig, *The Fundamentals of Density Functional Theory*. (EAG.LE., Leipzig, 2003).
[160] E. Engel and R. M. Dreizler, *Density Functional Theory: An Advanced Course*. (Springer, 2011).
[161] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis and J. A. Montgomery, J Comput Chem **14**, 1347 (1993).
[162] R. F. W. Bader, P. L. A. Popelier and T. A. Keith, Angew Chem **33**, 620 (1994).
[163] R. F. W. Bader and D. Bayles, J Phys Chem A **104**, 5579 (2000).
[164] J. S. Dehesa, S. Lopez-Rosa and D. Manzano, Eur Phys J D **55**, 539 (2009).
[165] L. Lorenzo and R. A. Mosquera, Chem Phys Lett **356**, 305 (2002).
[166] F. Cortes-Guzman and R. F. W. Bader, Chem Phys Lett **379**, 183 (2003).
[167] F. Cortes-Guzman and R. F. W. Bader, J Phys Org Chem **17**, 95 (2004).
[168] S. B. Liu, J Chem Phys **126**, 191107 (2007).
[169] C. Rong, T. Lu and S. Liu, J Chem Phys **140** (2014).
[170] M. H. Cohen, M. V. Gandugliapirovano and J. Kudrnovsky, J Chem Phys **103**, 3543 (1995).
[171] R. F. Nalewajski and E. Broniatowska, J Phys Chem A **107**, 6270 (2003).

[172] A. Luch, *The carcinogenic effects of polycyclic aromatic hydrocarbons*. (Imperial College Press ; Distributed by World Scientific Pub., London; Hackensack, NJ; London, 2005).
[173] R. S. Braga, P. M. V. B. Barone and D. S. Galvao, J Mol Struc-Theochem **464**, 257 (1999).
[174] N. Fjodorova, M. Vracko, M. Novic, A. Roncaglioni and E. Benfenati, Chem Cent J **4 Suppl 1**, S3 (2010).
[175] A. Pullman and B. Pullman, Adv Cancer Res **3**, 117 (1955).
[176] R. E. Lehr and D. M. Jerina, J Toxicol Environ Health **2**, 1259 (1977).
[177] K. P. Vijayalakshmi and C. H. Suresh, J Comput Chem **29**, 1808 (2008).
[178] K. Tanabe, B. Lucic, D. Amic, T. Kurita, M. Kaihara, N. Onodera and T. Suzuki, Mol Divers **14**, 789 (2010).
[179] O. Ivanciuc, Internet Electron. J. Mol. Des. **1**, 203 (2002).
[180] L. J. Marnett, Carcinogenesis **8**, 1365 (1987).
[181] F. P. Guengerich, J. A. Krauser and W. W. Johnson, Biochemistry **43**, 10775 (2004).
[182] A. H. Conney, Cancer Res **42**, 4875 (1982).
[183] M. Saeed, S. Higginbotham, N. Gaikwad, D. Chakravarti, E. Rogan and E. Cavalieri, Free Radic Biol Med **47**, 1075 (2009).
[184] J. A. Cohn, A. P. Alvares and A. Kappas, J Exp Med **145**, 1607 (1977).
[185] L. W. Wormhoudt, J. N. Commandeur and N. P. Vermeulen, Crit Rev Toxicol **29**, 59 (1999).

[186] J. Iball, Am J Cancer **35**, 188 (1939).

[187] L. Zhang, K. Sannes, A. J. Shusterman and C. Hansch, Chem Biol Interact **81**, 149 (1992).

[188] A. Leo, C. Hansch and D. Elkins, Chem Rev **71**, 525 (1971).

[189] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*. (John Wiley & Sons, 2008).

[190] J. Schuur and J. Gasteiger, Anal Chem **69**, 2398 (1997).

[191] V. Consonni, R. Todeschini, M. Pavan and P. Gramatica, J Chem Inf Comput Sci **42**, 693 (2002).

[192] J. Devillers and A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPAR*. (CRC Press, 2000).

[193] L. Breiman, Mach Learn **45**, 5 (2001).

[194] J. H. Friedman, Ann Stat **19**, 1 (1991).

[195] E. Estrada, J Chem Inf Comp Sci **37**, 320 (1997).

[196] C. E. Shannon, Bell Syst Tech J **27**, 379 (1948).

[197] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*. (John Wiley & Sons, 2009).

[198] A. Karatzoglou, T. U. t. Wien, A. Smola, K. Hornik and W. t. Wien, J Stat Softw, 1 (2004).

[199] R. D. C. Team, *{R: A language and environment for statistical computing}*. (R Foundation for Statistical Computing, Vienna, Austria, 2009).

[200] T. M. Therneau and E. J. Atkinson, *An introduction to recursive partitioning using the rpart routines. Divsion of Biostatistics 61*. (1997).

[201] W. N. Venables, B. D. Ripley and V. , W. N, *Modern applied statistics with S*. (Springer, New York, 2002).

[202] M. Culp, K. Johnson and G. Michailidis, J Stat Softw **17**, 1 (2006).

[203] R. Kohavi and F. Provost, Mach Learn **30**, 271 (1998).

[204] P. Zhang, Ann Statist **21**, 299 (1993).

[205] E. L. Cavalieri, E. G. Rogan, R. W. Roth, R. K. Saugier and A. Hakam, Chem-Biol Interact **47**, 87 (1983).

[206] L. Von Szentpaly, J. Am. Chem. Soc. **106**, 6021 (1984).

[207] G. P. Moss, Pure Appl Chem **70**, 143 (1998).

[208] E. Cavalieri and E. Rogan, Environ Health Perspect **64**, 69 (1985).

[209] *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. (National Institute of Standards and Technology, 2005).

[210] P. D. Devanesan, P. Cremonesi, J. E. Nunnally, E. G. Rogan and E. L. Cavalieri, Chem Res Toxicol **3**, 580 (1990).

[211] E. Clar and W. Schmidt, Tetrahedron **35**, 1027 (1979).

[212] S. Kullback and R. A. Leibler, Ann Math Statist **22**, 79 (1951).

[213] V. N. Vapnik, *The nature of statistical learning theory*. (Springer, New York, 1995).

[214] B. E. Boser, I. M. Guyon and V. N. Vapnik, 1992 (unpublished).

[215] V. N. Vapnik, *Statistical learning theory*. (Wiley, New York, 1998).

[216] T. Hastie, R. Tibshirani and J. H. Friedman, *The elements of statistical learning data mining, inference, and prediction*. (Springer, New York, 2009).

[217] C. Cortes and V. Vapnik, Mach Learn **20**, 273 (1995).

[218] B. Schölkopf, C. J. C. Burges and A. J. Smola, *Advances in Kernel Methods: Support Vector Learning*. (MIT Press, 1999).

[219] V. N. Vapnik, IEEE Trans Neural Netw 10, 988 (1999).

[220] Schölkopf, Bernhard, C. J. C. Burges and A. J. Smola, *Advances in Kernel Methods: Support Vector Learning*. (MIT Press, 1999).