NTNU

Norwegian University of
Science and Technology

Predicting the singlet-triplet energy gap of blue organic light emitting diodes

Signe Eva Helmersen

Chemistry
Submission date: June 2018
Supervisor: Per-Olof Åstrand, IKJ
Co-supervisor: Vishwesh Venkatraman, IKJ

Norwegian University of Science and Technology
Department of Chemistry

Acknowledgment

The work presented in this thesis has been performed at the Department of Chemistry at the Norwegian University of Science and Technology (NTNU), Trondheim. It has been conducted during the time period February to June 2018.

I would first like to thank professor Bjørn Kåre Alsberg for welcoming me with open arm as a part of his research team. He truly believed in me, supported me and motivated me to go after what I believe is interesting and not what is easy.

I am grateful to my supervisor Per-Olof Åstrand for taking me on as his master student given the circumstances and for allowing me to continue the work initially started during the master project conducted the fall of 2017.

An last but not least, I want to thank my co-supervisor Vishwesh Venkatraman. He has truly been my go-to expert. He has provided a tremendous amount of help ranging from how to write the thesis to programming problems. He has a unique ability to explain complex theory and problems in a way such that a novice like me could understand and learn more.

To my family and friend, thanks for the encouragement, the support and company during ups and downs. And a special thanks to my sister for reviewing my thesis, being a shoulder to cry on when thing been tough and generally being there when I needed it the most.

Abstract

In order to speed up molecular design efforts, we examine the utility of chemometric approaches to estimate ΔE_{st} of blue TADF based OLEDs. In order to create efficient blue emitting OLEDs, ΔE_{st} is required to be sufficiently small. This makes it an import part of the design strategy for blue OLED. To this end structural and experimental data for 60 different blue emitting dyes were collected from a recent review paper. Various 3D molecular descriptors based on energies, charges and geometry were calculated using KRAKENX and used to identify quantitative structure-property relationships (QSPR) for ΔE_{st}. Exploratory analysis using principal component analysis (PCA) and k-means cluster analysis were performed to observe the variance in the data and to detect patterns and potential outliers. Partial least squares regression (PLSR) and the non-linear regression tree method Cubist was used to create models for estimating ΔE_{st}. Based on the result obtained from the exploratory analysis, different approaches were tested to potentially improve the models. All the models created in this thesis were incapable of producing high accuracy estimates of ΔE_{st}. The best results were obtained with Cubist, indicating that there is a non-linear relation between the descriptors and the ΔE_{st} property. The best model had a R^2_{CV} of 0.59 and a RMSE of 0.08 with variable selection performed. The data used in this thesis proved to be highly heterogeneous, which resulted in a insufficient coverage of the chemical space. Also, the descriptors chosen may not be the most ideal to obtain QSPR for ΔE_{st}. An attempt was made to calculate ΔE_{st} using TDDFT on a selected set of structures. The computation was unfortunately not completed due to time constraints and problems with convergence.

Sammendrag

I et forsøk på å assistere design av nye blå OLEDs har vi i denne oppgave forsøkt å utnytte kjemometri til å estimere ΔE_{st}. For å kunne oppnå effektive OLEDs basert på TADF må ΔE_{st} være tilstrekkelig liten. ΔE_{st} representerer derfor en kritisk parameter i design av nye blå organiske fargestoff. Strukturdata og eksperimentelle verdier for 60 ulike blå fargestoff ble hentet fra en nylig utgitt oversiktsartikkel. Ulike 3D molekylære deskriptorer ble generert ved bruk av programvaren KRAKENX for å identifisere QSPR for ΔE_{st}. *Principal component analysis* (PCA) og *k-means cluster analysis* ble gjennomført for å vurdere variansen i dataene og for å kunne identifisere mønstre og potensielle punkt liggende utenfor det kjemiske rommet. Lineær *partial least squares regression* (PLSR) og den ikke-lineær metoden Cubist ble brukt for å generere ulike modeller for å estimere ΔE_{st}. Ingen av modellene produsert i denne oppgaven viser tilstrekkelig nøyaktighet i estimeringen av ΔE_{st}. Den beste modellen ble oppnådd ved bruk av Cubist metoden, som indikerer at det er en ikke-lineær relasjon mellom deskriptorene of ΔE_{st}. Denne modellen hadde en R^2_{CV} verdi på 0.59 og en RMSE på 0.08 når variabel seleksjon ble utført. Resultatene viser at dataene brukt i denne oppgaven er svært heterogene. De molekylære deskriptorene brukt i denne oppgaven representerer også et usikkerhetsmoment. Det er vanskelig å vite nøyaktig hvilke deskriptorer som er best egnet for å danne QSPR for ΔE_{st}. Det er derfor ikke sikkert at de deskriptorene som ble valgt i denne oppgaven er de som er mest relevante for å kunne estimere ΔE_{st}. Det ble også forsøkt gjennomført en TDDFT beregning for utvalgte strukturer for å bestemme ΔE_{st}. Dessverre lyktes vi ikke med å fullføre disse beregningene innenfor den aktuelle tidsrammen grunnet problemer knyttet til konvergens.

Contents

1 Introduction

Ever since the invention of the first Organic Light Emitting Diode (OLED) in 1987 by Tang and Van Slyke [1], much research has been devoted to the subject as it represents great advantages in both lighting and display technologies [2]. Some of these advantages are improved image quality and contrast, faster response time, faster refresh rate, wider viewing angles and thinner and lighter devices. Thus, the application range of such OLEDs are extensive and some of them are presented in Fig. 1.1.

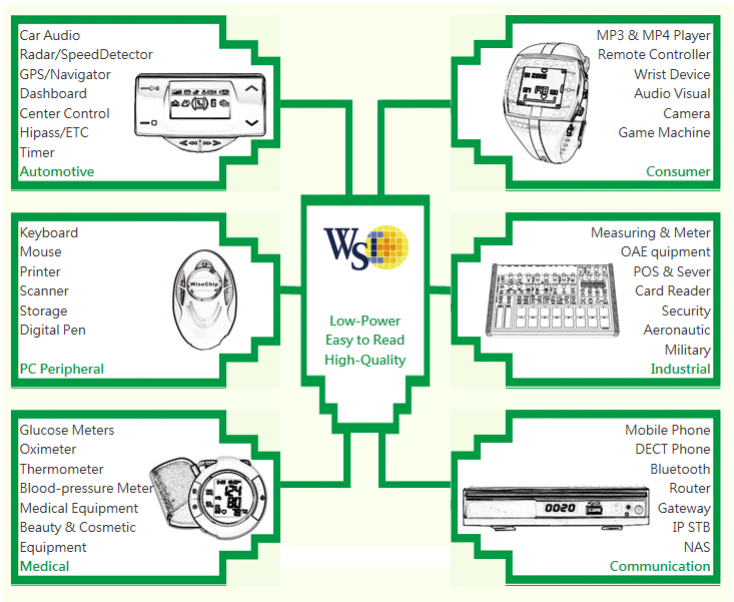
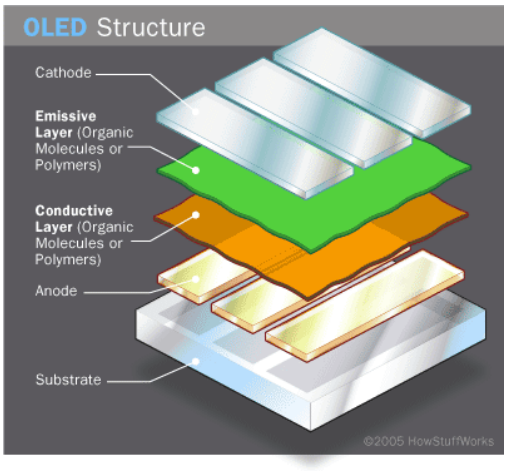


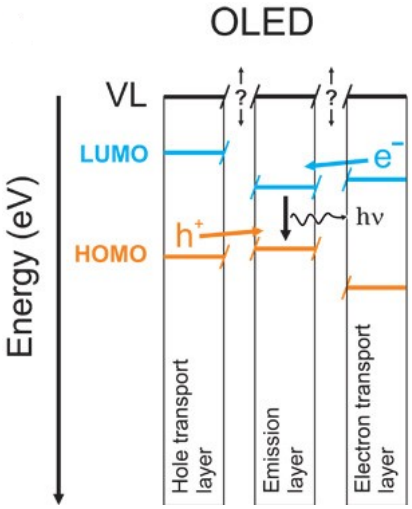
Figure 1.1: Application possibilities for OLEDs. The figure is taken from `http://www.braemac.co.uk/oledapplications.html` [3]

For the OLEDs to be applicable for the commercial market there are some challenges that have to be addressed. The devices should have high photoluminescence quantum yield, in particular a high external quantum efficiency [2, 4]. The organic material used in the OLED must also be morphologically stable and demonstrate thermal stability. The energy levels of the frontier orbitals, i.e. highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), in each layer of the device should be reasonably aligned. The alignment of the en-

ergy levels is important for the control of the exciton recombination region and to minimize the barrier to charge injection [2, 5]. A simplified representation of an OLED cell is presented in Fig. 1.2a and the energy alignment is presented in Fig. 1.2b.



(a) A simplified representation of an OLED cell where an organic emissive layer and a organic conducting layer is situated between the electrodes. The figure is taken from website `https://electronics.howstuffworks.com/oled1.htm` written by Craig Freudenrich [6]



(b) Example of energy alignment of HOMO and LUMO in the different layers of an OLED cell. The figure is taken from an article by Martin Oehzelt et al. [7]

Figure 1.2: A simplified representation of the layers in an OLED cell and an example of the energy level alignment in those layers

Another key feature that is of great importance is the management of hole and electron recombination[2]. Exciton formation through charge recombination results in 25 % singlet and 75 % triplet excited states. Only the former contributes to light emission by fluorescence whilst the latter contributes by phosphorescent emission. However, the triplet excitons can contribute to fluorescent emission by means of inter system crossing (ISC) and reversed inter system crossing (RISC)

[8]. In order for RISC to occur, the energy difference between the lowest singlet and triplet excited state, termed ΔE_{st}, must be sufficiently small. Generally the triplet state lies lower in energy than the singlet state, which indicates that the transition from triplet to singlet states must be thermally activated by the thermal motion of the organic molecules. A low ΔE_{st} makes a thermal upconversion from the triplet state to the singlet state possible. Although much research has been devoted to OLEDs there are still problems related to the construction of high performance blue TADF based OLEDs [9, 10, 11].

1.1 Organic Light Emitting Diodes

Most organic semiconductors used in OLEDs are mainly amorphous π-conjugated system and thus differ from the conventional valence band theory. The conjugation of the systems determines the conductivity of the the organic semiconductor. The energy levels, in particular the energy band gap of the emitting material, determines the wavelength of the light emitted. A suitable energy level alignment of the different layers in the diode is important in order to minimize the barrier to charge injection and to control the recombination region within the device. Also, the alignment is important because of its great impact on the external quantum efficiency, luminance and lifetime of the OLEDs[5, 2].

OLEDs normally consist of two layers of organic materials combined with a cathode and an anode. These two organic layers represents the conductive and the emissive part of the diode. Electrons and holes are injected into the organic layers on either side of the diode through the electrodes creating self-localized electronic states such as excitons [12, 13]. In a semiconducting organic material an exciton is a bound electron-hole pair state formed due to strong electron-lattice interactions. The efficiency of the OLED is related to the exciton dissociation mechanism [13, 14]

When the electron and hole recombination occurs the resulting excited state can either be a singlet state or an triplet state due to the spin wavefunction formed from the two spin electronic charges[8]. According to spin statistics, 25 % of the excited states produced through exciton formation corresponds to singlet states and 75 % triplet states [2]. Since only singlets fluoresce, and given that the exchange energy

is large, cross-over from triplet to singlet is highly unlikely to happen [14]. But, triplet states can contribute to the recombination radiation through either phosphoroscence, thermally activated delayed fluorescence (TADF) or triplet-triplet annihilation [8]. A device of solely fluorescence light emission can only yield a theoretical external quantum efficiency (EQE) of 25 %, but by exploiting both the singlet and triplet states, an internal quantum efficiency (IQE) of 100 % is achievable [2, 8].

The first OLED based on pure organic TADF emitter was reported in 2011 by Adachi *et al.* [15] and has ever since been a subject of tremendous research in order to improve the device's performance [16]. In thermally activated delayed fluorescent OLEDs a charge transfer (CT) from excited singlet state to excited triplet state occurs through an intersystem crossing (ISC) followed by a reversed intersystem crossing (RISC) and relaxation to ground state [4]. See Fig. 1.3 illustrating the different transitions. According to quantum mechanical theory of selection rule, this ISC and RISC is spin forbidden as charge transfer between states with different spin multiplicities is theoretically not allowed [17]. TADF yields an emission with a longer lifetime than the direct fluorescence since the process of ISC and RISC are slow processes due to the change of electron spin between the states [4]. The difference between the lowest excited singlet state and lowest excited triplet state (S_1 and T_1 in Fig. 1.3 respectively) is denoted ΔE_{st}. Usually T_1 lies lower in energy than S_1 meaning that the transition from T_1 to S_1 is an endothermic process. The ΔE_{st} is thus desired to be small in order to thermally activate the process. The RISC can be thermally activated by thermal motion of the organic molecule at sufficient temperatures (> 300 K)[4].

Adatchi *et al.* showed that by reducing the overlap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), the ΔE_{st} could be reduced yielding more efficient TADF OLEDs [15]. This allows for all the excited states to be harvested, both singlet and triplet, resulting in an increase in the exciton statistical limit of fluorescent materials from 25 % to 100 %. The relation between ΔE_{st}, RISC and temperature can be expressed through

following Boltzmann distribution,

$$k_{RISC} \propto \exp\left(\frac{\Delta E_{st}}{k_B T}\right) \tag{1.1}$$

where k_{RISK} is the rate constant of RISC, k_B is the Boltzmann's constant and T is the temperature [2, 4]. Also, ΔE_{st} is related to the structure of the emitter through the relation,

$$\Delta E_{st} = E_S - E_T = 2J \tag{1.2}$$

where E_S and E_T is the energy of the singlet and triplet state respectively and J is the exchange integral. J depends on the overlap between HOMO and LUMO, i.e. small overlap yields small ΔE_{st}. In addition to ΔE_{st} and k_{RISC}, the radiative rate constant, r_r, of the singlet exciton transition from S_1 to S_0 is important to get efficient TADF emission.

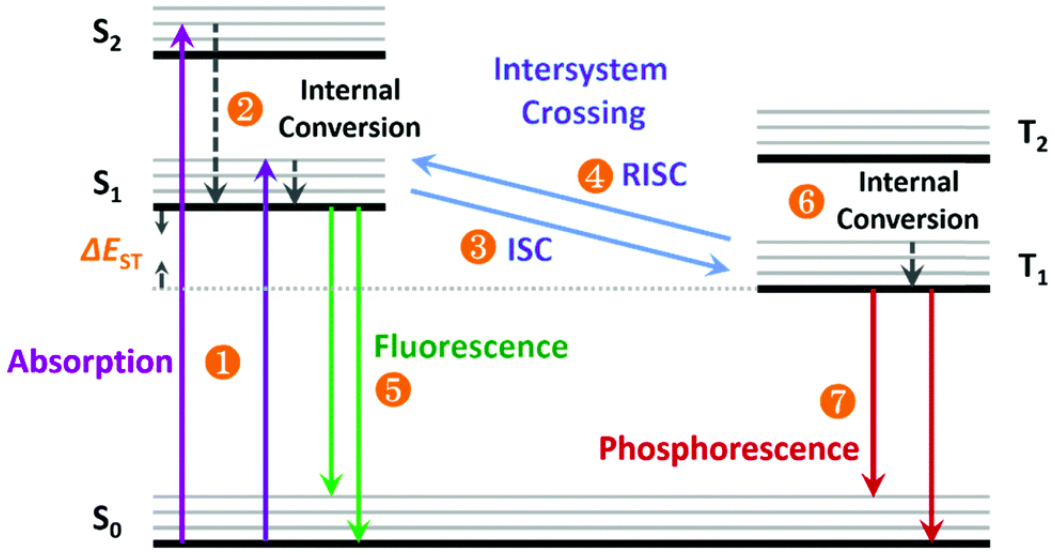


Figure 1.3: The transition processes in a TADF based OLED follow the steps 1, 2, 3, 4 and 5. The figure is taken from an article by Yang et al. [4]

2 Methodology

Quantum chemical calculations and experimental procedures can be used to determine ΔE_{st} for a given structure. However, these methods are very time consuming. The purpose of the study presented here is to used chemometrics in an attempt to create regression models which can yield reliable ΔE_{st} responses based on calculated molecular descriptors. If such models could be constructed it would be a great advantage because it would reduce the time required to determine ΔE_{st}. To determine ΔE_{st} of an unknown sample one would only need to calculate the molecular descriptors for the sample and employ them to the model. The reason for choosing ΔE_{st} and not some other property is because it is a critical property of blue TADF based OLEDs. As presented in the introduction the ΔE_{st} must be sufficiently small in order to achieve TADF. Also, experimental values are often reported in articles on OLED dyes, which is needed in the regression model. To the best of our knowledge ΔE_{st} has not been used as the property of interest in a QSPR model before. Nantasenamat et al. and Chen et al. have both done QSPR studies on the emission maximum of different dyes [18, 19]. Barbosa et al. have studied the glass transition temperature of different OLED materials [20]. The estimation of ΔE_{st} thus represent uncharted territory. The most relevant molecular representations for describing ΔE_{st} is thus unknown. In this study some well known and frequently used descriptors are calculated as a starting point.

In this study, experimental values of ΔE_{st} for blue light emitting dyes have been collected from an resent review paper [2]. Different 3D molecular descriptors based on energies, charges and geometry were used to identify quantitative structure-property relationships (QSPR) for ΔE_{st}. The exploratory analysis methods principal component analysis (PCA) and k-means cluster analysis were performed on the descriptors to study the distribution of the data and to evaluate patterns and potential outliers. Regression analysis was then performed using the experimental data and the molecular descriptors. The linear regression method partial least squares regression (PLSR) and the non-linear regression tree method Cubist where chosen in this study because they are easily interpreted. In order to validate some of the experimentally determined ΔE_{st} values and responses from the regression

model, density functional theory (DFT) and time-dependent density functional theory (TDDFT) was performed on selected structures.

The work presented in this thesis is a continuation of the work done on the master project conducted fall of 2017. The result of the PCA and PLSR calculations where the full data set is used was obtained during that project. However, the discussion of of the result has been altered. In order to avoid confusion, results obtained during the project will be specified in the relating plots and figures.

Chemometrics has for over 30 years been developing classification and regression methods able to provide reliable models, both for reproducing experimental data and for predicting unknown values. The use and interest for reliable prediction models has been growing in the last few years as they are considered to be useful and safer tools for predicting data for chemicals. Chemoinformatics encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information [21, 22]. Exploratory data analysis consist of techniques used on a data set to disclose information about the data distribution, outliers, clusters and relationship between objects and/or variables. Principal component analysis (PCA) and k-means cluster analysis are two examples of such techniques, which are used in this study. PCA and k-means cluster analysis will be explained in more detail in Section 3.1 and Section 3.4, respectively. Another topic within chemometrics is regression analysis where the purpose is to find a relationship between experimental data and molecular descriptors. The regression model that describe this relationship can then be used to predict future unknown samples. Two different regression methods which are used in this study are partial least squares regression (PLSR) and Cubist, which will be explained in more detail in Section 3.2 and Section 3.3, respectively. Chemometrics provides useful tools for data analysis and modelling in quantitative-structure property relationship (QSPR) methods. The development of a QSPR model requires three fundamental components: i) experimental data of a biological activity or property for a group of chemicals, ii) molecular descriptors and iii) mathematical methods to find the relationships between a molecular property and the molecular structure [23]. The accuracy of a property estimated by the use of QSPR methods

are dependent on the accuracy of the input data to the model. It is therefore of vital importance to obtain high-accuracy experimental data and relevant molecular representation of the structures used in a QSPR model in order to achieve an accurate predictive model. The molecular representations of the structures should be chosen such that they reflect the property of interest. This is not an easy task, because this is not necessarily known *a priori*.

The work done is this study is structured according to the development of a QSPR model described in the previous section. A simple representation of the work flow is depicted in Fig. 2.1. All the calculations, except the DFT and the TDDFT computations, were done on a desktop computer with Intel Core i5-7200U, 2,5 GHz and 8 GB RAM.

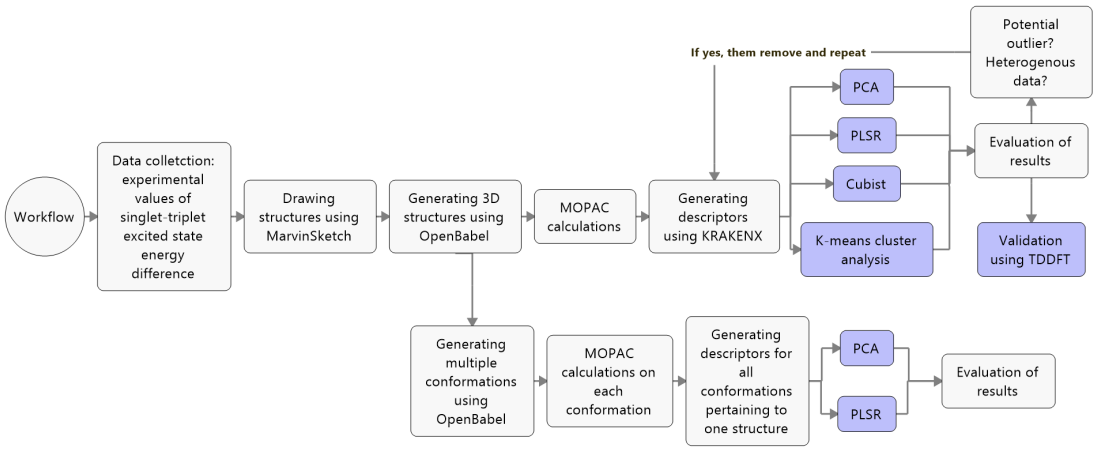


Figure 2.1: A simple representation of the work flow.

2.1 Data collection

Experimental and structure data for blue TADF based OLEDs were collated from a recent review article [2], and is presented in Table 2.1. All of the structures was drawn in two dimensions using MarvinSketch [24] and is attached in Appendix A. A three dimensional representation of all the structures were generated using Open Babel [25].

Table 2.1: Experimentally determined values of ΔE_{st}.

Structure name	ΔE_{st}	ref.
1	0.54	[26]
2	0.45	[26]
3	0.32	[26]
2DAC-Mes3B	0.058	[27]
2PXZ-TAZ	0.23	[28]
34TCzPN	0.16	[29]
35IPNDCz	0.14	[30]
3CzFCN	0.06	[31]
44TCzPN	0.21	[29]
4CzFCN	0.06	[31]
5CzCF3Ph	0.02	[32]
Ac-HPM	0.18	[33]
Ac-MPM	0.19	[33]
Ac-PPM	0.19	[33]
ACRPOB	0.06–0.12	[34]
ACRSA	0.03	[35]
ATP-ACR	0.16	[36]
BCzT	0.29–0.33	[37]
BFCz-2CN	0.13	[38]
BTCz-2CN	0.17	[38]
CC2BP	0.14	[39]
CCT2A	0.06	[40]
CNBPCz	0.27	[41]
CPC	0.04	[42]
Cz2BP	0.21	[39]
CzAcSF	0.14	[43]
CzBPCN	0.27	[41]
DABNA-1	0.18	[44]
DABNA-2	0.14	[44]

DAC-BTZ	0.18–0.22	[45]
DAC-Mes3B	0.062	[27]
DCBPy	0.07	[46]
DCN-3	0.13	[47]
DCzIPN	0.05	[48]
DCzmCzTrz	0.20	[49]
DCzTrz	0.25	[50]
DDCzIPN	0.13	[51]
DDCzTrz	0.27	[50]
DMAC-DPS	0.08	[52]
DMAC-PXB	0.013	[53]
DMAC-TRZ	0.046	[54]
DMOC-DPS	0.21	[55]
DPXZPO	0.19	[56]
DPAA-AF	0.021	[57]
DTC-mBPSB	0.24	[58]
DTC-pBPSB	0.05	[58]
DTPDDA	0.14	[59]
m-ATP-ACR	0.13	[60, 36]
m-ATP-CDP	0.26	[60, 36]
mPTC	0.01	[61]
PPZ–4TPT	0.43	[52]
SFDPAPOB	0.06–0.12	[34]
SpiroAC-TRZ	0.072	[62]
SPXZPO	0.26	[56]
SXDPAPOB	0.06–0.12	[34]
TB-1PXZ	0.12	[63]
TB-2PXZ	0.05	[63]
TCzTrz	0.16	[49]
TMCPOB	0.06–0.12	[34]
TPXZPO	0.11	[56]

Studying all the different structures in the data set it is evident that they consist of a multitude of different combinations of donor and acceptor units. The architecture is also quite different, including donor-acceptor (D-A), donor-acceptor-donor (D-A-D), combination of two D-A-D, acceptor-$n$donors (A-nD) and donor-acceptor-acceptor-donor(D-A-A-D). These different designs are proposed to spatially separate HOMO and LUMO in order to reduce the overlap. As mentioned earlier, a reduced overlap has shown to decrease the ΔE_{st}. There is one exception to this type of design strategy among the structures. DABNA-1 and DABNA-2 have HOMO and LUMO separation as a result of multiple resonance effects, not as a result of spatial separation[44]. Regardless, all structures have been reported to have sufficiently small HOMO-LUMO overlap which yield small ΔE_{st}. The structures are also quite different in size, with the largest consisting of 126 atoms and the smallest consisting of 54 atoms. A first impression of the structures is that they look fairly rigid. As a consequence it is expected that there will be few or no multiple conformations available.

2.1.1 Molecular representation

Molecular representations are descriptions of a molecular system. From these molecular representations different molecular descriptors, which yield different types of chemical information, can be computed. Descriptors, which are quantitative descriptions of molecular structures, can be computed as a combination of atomic properties and the distribution of these properties in the molecular structure [64]. Descriptors can be grouped into five categories

- Descriptors that can be derived from molecular formulas, such as molecular weight
- Descriptors that depend on constitutions such as topological surface area
- Configuration dependent descriptors
- Conformation dependent descriptors
- Descriptors that take conformational flexibility into account

The most useful descriptors are those with the highest information density and minimal addition of noise. Descriptors play a fundamental role in chemistry, biology and many other fields, and are as mentioned of great importance in research fields of QSPR [65]. There are more than 5000 descriptors derived from different theories and approaches up til today [66], and most of them can be computed by means of software applications[65]. For a detailed description of many different descriptors the reader is referred to the book *Molecular Descriptors for Chemoinformatics Volume I: Alphabetical Listing / Volume II: Appendices, References* by R. Todeschini and V. Consonni [23].

Molecular properties was calculated for each structure using the Molecular Orbital PACkage (MOPAC), which is a semi-empirical quantum chemistry program based on Dewar and Thiel's neglect of diatomic differential approximation (NDDO) [67]. The program uses concepts of quantum theory and thermodynamics to calculate molecular properties such as molecular orbital energies, heat of formation, electrophilic and nucleophilic delocalizability, etc. KRAKENX was used to calculate different molecular descriptors. KRAKENX is an open-source software [68] that calculates a variety of different groups of descriptors: vibrational frequency based eigenvalue descriptors (EVA), molecular orbital energy based electronic eigenvalue descriptors (EEVA), charge partial surface area (CPSA), Weighted Holistic Invariant Molecular (WHIM), BCUT, radial distribution function (RDF), autocorrelation, 3D molecule representation of structures based on electron diffraction (3D-MORSE), graph eigenvalues, geometry, charge and MOPAC generated energies [69]. For the purpose of predicting ΔE_{st} the CPSA, autocorrelation, charge, geometry, EVA, EEVA and MOPAC descriptors were calculated. Here the MOPAC descriptors refer to the molecular properties which was calculated using MOPAC. Some of the descriptors used in this study are explained in more detail in the following paragraphs.

The ovality index, O, is an anisometry (having unsymmetrical parts) descriptor based on the property that the spherical shape presents the minimum surface[70, 71]. It is calculated from the ratio between the actual molecular surface area (SA) and the minimum surface area (SA_0), corresponding to the Van der Waals volume (V_{vdw})

$$O = \frac{SA}{SA_0} = \frac{SA}{4\pi R^2} = \frac{SA}{4\pi(\frac{3V_{vdw}}{4\pi})^{2/3}} \tag{2.1}$$

The ovality index is equal to 1 for spherical molecules and increase with increasing linearity of the molecule

Structure 1 has the largest ovality factor, $O = 38$, and is depicted in Fig. 2.2. The structure with the lowest ovality factor, $O = 2.39$, is attributed to DABNA-1 and is depicted in Fig. 2.3.

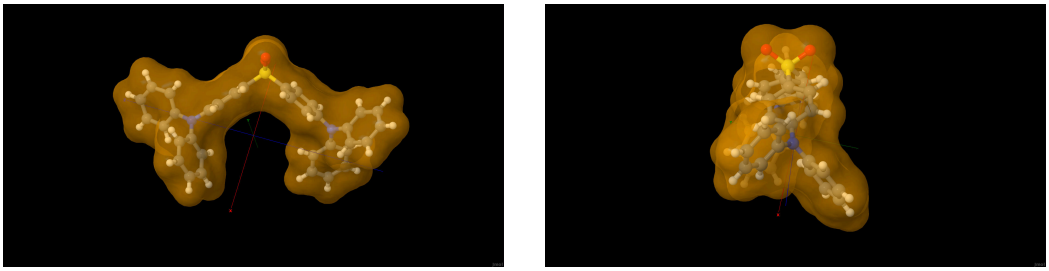


Figure 2.2: Molecular surface of structure 1 which displays the largest ovality with $O = 38$

Figure 2.3: Molecular surface of structure DABNA-1 which displays the lowest ovality with $O = 2.39$

The inverse of the ovality index is the globularity factor, G, which is between zero and one[72]. The most spherical molecules have globularity factors approximating unity, but for two molecules that have non-equal volumes the globularity factor reflects the relative compactness. When both the effective surface area and the volume of the molecule are available, the surface-volume ratio $G' = SA/V$ can be interpreted as a measure of the capability of a compound to adapt its shape to the requirements of an approaching reagent[73].

Fig. 2.4 and Fig. 2.5 shows the molecular surface of structure DDCzIPN and CC2TA, respectively. DDCzIPN has the largest globularity factor $G = 0.81$ and CC2TA the smallest with $G = 0.03$

Figure 2.4: Molecular surface of structure DDCzIPN which displays the largest globularity with $G = 0.81$

Figure 2.5: Molecular surface of structure CC2TA which displays the lowest globularity with $G = 0.03$

Asphericity, Ω_A, is also a shape descriptor which measures the deviations from a spherical shape[74]. It is calculated from the eigenvalues, λ_i, of the inertia matrix and takes values between 0 and 1.

$$\Omega_A = \frac{1}{2}\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2}{\lambda_1^2 + \lambda_2^2 + \lambda_3^2} \tag{2.2}$$

The Molecular eccentricity (ε) is a shape descriptor defined as

$$\varepsilon = \frac{(I_A^2 - I_C^2)^{1/2}}{I_A} \tag{2.3}$$

where I is the principal moment of inertia, which is a physical quantity related to the rotational dynamics of a molecule [74]. The subscripts A, B and C label the principal inertia axes. The molecular eccentricity takes values between zero and one where $\varepsilon = 0$ corresponds to a spherical molecule and $\varepsilon = 1$ to a linear molecule.

Fig. 2.6 pictures the molecular surface of structure CC2BP which has the largest molecular eccentricity with $\varepsilon = 0.994$. The structure with the lowest molecular eccentricity $\varepsilon = 0.54$ is DDCzIPN, pictured in Fig. 2.7.

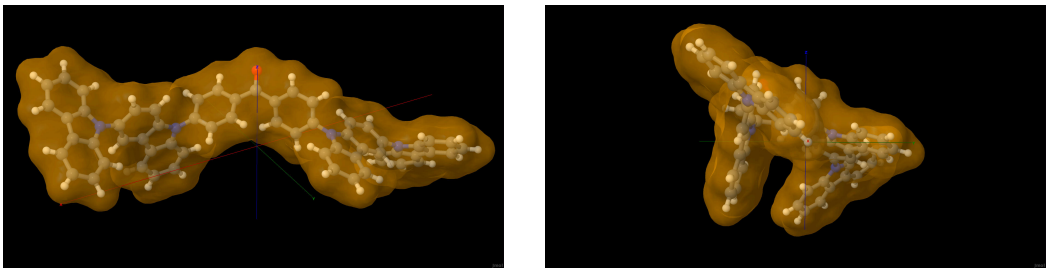


Figure 2.6: Molecular surface of structure CC2BP which displays the largest molecular eccentricity with $\varepsilon = 0.994$

Figure 2.7: Molecular surface of structure DDCzIPN which displays the lowest molecular eccentricity with $\varepsilon = 0.54$

The CPSA descriptors combine shape and electronic information to characterize molecules, thus they encode features responsible for polar interactions between molecules [75]. In deriving the CPSA descriptors the molecules are viewed as hard spheres defined by the Van der Waals radius and the surface area used is the solvent-accessible surface area. The contact surface where polar interactions can take place is characterized by a specific electronic distribution obtained by mapping atomic partial charges on the solvent-accessible surface. Some of the CPSA descriptors are defined as [23]:

- PNSA-1: partial negative surface area, the sum of the solvent-accessible surface area of all negatively charged atoms
- DPSA-3: difference in atomic charge weighted surface area, the atomic charge weighted positive solvent-accessible surface area minus the atomic charge weighted negative solvent-accessible surface area (DPSA-3=PPSA-3-PNSA-3)
- PPSA-3: atomic charge weighted positive surface area is the sum of the product of atomic solvent-accessible surface area and the partial charge over all positively charged atoms
- PNSA-3: atomic charge weighted negative surface area is the sum of the product of atomic solvent-accessible surface area and the partial charge over all negatively charged atoms

- FNSA-2: fractional charged partial negative surface area, the total charge weighted negative surface area divided by the total molecular solvent-accessible surface area (SASA)
- PPSA-5: PPSA-1 multiplied with the sum of all positive charges, divided by the number of positively charged atoms [76]
- PNSA-4: PNSA-1 multiplied with the total negative charge, divided by the total number of atoms[76]
- PNSA-5: PNSA-1 multiplied with the total negative charge, divided by the total number of negatively charged atoms
- RNCS: relative negative charge surface area, is the solvent-accessible surface area of the most negative atom divided by the relative negative charge (RNCG)
- RNCG: relative negative charge is the partial charge of the most negative atom divided by the total negative charge

3-dimensional autocorrelation (3DA) generates a histogram of atom pair distances within a molecule up to a cutoff distance. Interatomic distances are represented in terms of Euclidian distances. To extend the 3DA descriptor beyond the geometric characteristics of a molecule, atom pair distances are weighted by atom properties. The formal definition of 3DA is

$$\text{Autocorrelation}(r_a, r_b) = \sum_{i}^{n} \sum_{j}^{n} \delta(r_a \leq r_{ij} < r_b) P_i P_j \tag{2.4}$$

where r_{ij} is the distance between atom i and j, and n is the total number of atoms in the molecule. r_a and r_b represent the lower and upper cutoff distance, respectively. P_i and P_j are the atomic properties for atom i and j used to weight the autocorrelation. Weighting of the 3DA allows the descriptor to explain the distribution of specific atom properties within a molecule. The problem arises when the property used to weight the atom distances are heterogeneously signed, e.g

partial charge. Significant information loss arises due to sign-cancellation. Gregory Sliwoski et al. propose a solution by separating the 3DA histogram into three parts, negative-negative, positive-positive and opposite sign property pairs. By doing this the different types of information loss are revealed. 3D descriptors such as 3DA are computed from a single structure conformation. As interatomic distances increase, the degree of flexibility and rotatable bonds may increase, leading to greater degrees of conformational uncertainty at larger distances. In this study the autocorrelation is weighted by charges. In order to distinguish between the different autocorrelation descriptors they have been given a number, e.g. "autocorrelated charge property 1".

MOPAC descriptors are mostly energy descriptors including HOMO and LUMO energies, HOMO-LUMO fraction, total dipole moment, maximum and minimum electrophilic delocalizability (DER) and nucleophilic delocalizability (DNR), among others. The HOMO-LUMO fraction is simply the ratio of HOMO energies to LUMO energies. The electrophilic delocalizability is defined as [77]

$$D^N(r) = 2\sum_{k}^{vac}\sum_{\mu(r)}\frac{c_{\mu k}^2}{\alpha - \epsilon_k} \tag{2.5}$$

and the nucleophilic delocalizability is defined as

$$D^E(r) = 2\sum_{i}^{occ}\sum_{\mu(r)}\frac{c_{\mu i}^2}{\epsilon_i - \alpha} \tag{2.6}$$

The outer sum in Eq. (2.5) and Eq. (2.6) goes over all vacant MOs, k, and all occupied MOs, i, respectively. The inner sum put together the contributions of all atomic orbitals (AO), μ, belonging to the reagent center, r, of interest. $c_{\mu j}$ is the linear combination of atomic orbitals (LCAO) of AO, μ, at a reagent center, r, of MO, j. ϵ_j, is the energy of the jth MO. α is defined as the average of the HOMO and LUMO energies, i.e.

$$\alpha = \frac{1}{2}(\epsilon_{HOMO} + \epsilon_{LUMO}) \tag{2.7}$$

The delocalizability was introduced by Fukui et al. [78] as a reactivity index for saturated compounds, thus generalizing the superdelocalizability of conjugated compounds[79].

Fig. 2.8 shows the electrostatic potential surface for the structures that displayed the largest and smallest value of total dipole moment. DTC-mBPSB has a total dipole of 9.72 Debye and represents the structure with the largest total dipole. The smallest total dipole is attributed to the DDCzIPN structure with a value of 0.05 Debye.

Figure 2.8: Electrostatic potential surface of structure DTC-mBPSB (left) and DDCzIPN (right), which displays the largest total dipole with a value of 9.72 Debye and the lowest total dipole with a value of 0.05 Debye, respectively.

Vibrational frequency based eigenvalue descriptors (EVA) developed by Ferguson et al. is based on molecular vibrational motion [80, 81]. The idea is that, given the implicit dependence of the vibrational frequencies and the atomic displacement on the molecular wave function, this descriptor would yield reliable characterization of the molecular structure with adequate information on the shape, size and electronic properties of the molecule. The vibrational frequencies are projected onto a bounded frequency scale with individual vibrations represented along the axis. For a system of N atoms there are 3N-6 normal modes of vibration, all having a unique frequency of vibration. Each vibration is then represented by a Gaussian curve which is dependent of the vibrational frequency and the standard deviation, σ. The resultant spectrum is then sampled at fixed increments. At each incremental point a EVA value is calculated as the sum of amplitudes of the overlapping

Gaussian function. The typical range of the bounded frequency scale is from 0 to 4000 cm^{-1} to cover all fundamental vibrations. Typical σ value is 10 cm^{-1} and increment size is 5 cm^{-1}. The EVA descriptors calculated in this study were calculated using a frequency scare ranging from 1-4000 cm^{-1}, sigma value of 2 cm^{-1} and an increment size of 1 cm^{-1}.

Molecular orbital energy based electronic eigenvalue descriptors (EEVA) is a modification of the EVA descriptor[82]. The same principals are used, but with the molecular orbital (MO) energies rather than the vibrational frequencies. The MO energies are projected onto a bounded energy scale and a Gaussian function is placed over each eigenvalue. As for the EVA descriptor, at given increments the EEVA is calculated as the sum of the overlapping Gaussian functions. Here the energy scale is usually ranging from -45 to 10 eV. The σ value is often set to 0.5 eV and the size of the increments to 0.25 eV. The EEVA descriptors used in this study is calculated with the energy scale ranging from -45 to 10 eV, a sigma value of 0.05 eV and an increment size of 0.025 eV.

2.2 Exploratory analysis

PCA were performed using an in-house GUI based application. The statistical computer software R [83] was used to preprocess the descriptor data before subjecting them to PCA. The preprocessing consisted of removing all "Not available" (Na) entries and near zero variance columns. The cleaned descriptor data was then incrementally subjected to PCA with different groups of descriptors, i.e. the PCA was run for EVA, EEVA, EVA and EEVA, REST, EVA and REST, EEVA and REST and finally for all the descriptors. Here the REST abbreviation means all descriptors except for EVA and EEVA, i.e. autocorrelation, CPSA, MOPAC, charge and geometry descriptors. Appropriate plots were generated in order to evaluate the results.

In order to investigate how multiple structural conformations would impact the model performance, multiple conformations were generated using the confab package in Open Babel[25]. The maximum number of confirmations to be generated was set to ten. The default RMSD cutoff was 0.5 Å and the default energy cutoff

was 50.0 kcal/mol. Given these cutoffs there were only five structures that had multiple confirmations, namely DMAC-DPS, DMAC-TRZ, DTPDDA, SpiroAC-TRZ and SPXZPO. The descriptors for the multiple conformations pertaining to one structure were then calculated collectively using KRAKENX. Two different strategies were used to process the descriptors of the multiple conformations. One strategy was to take the mean of the descriptors. The other method was to use the Boltzmann weights of each conformation. After either of these methods, the descriptors was added to the descriptors for the rest of the data set. Then, PLSR were performed according to the procedure presented in Section 2.3.

K-means cluster analysis was performed and visualized utilizing the `NbClust` package [84] and `factoextra` package [85] in R. The optimal number of clusters was determined using the method of gap statistics. The descriptor input was scaled and the Na entries removed before employing them to the `kmeans` function.

2.3 Regression analysis

The PLSR modelling were performed using a PLSR script retrieved from `kraken-miner.com` [68]. The script utilizes the R packages `pls` [86], `parallel` [87], `plyr` [88] and `getopt` [89]. Inputs given to the script were the preprocessed descriptors and the experimental ΔE_{st} values. The third root of the ΔE_{st} values was used in order to get an approximate normal distribution. Note that some of the reported experimental values are given in an interval, for which the mean value have been used instead. In addition to the input, several conditions were set. 25 % of the calibration data were ascribed as test set. A random seed value was given and used for all the calculations to ensure the same objects in the test set for each calculation. Calculations included randomization test of the dependent variables, 10-fold cross-validation and VIP based variable selection. The VIP selection criteria ranged from a VIP value of 0.8 to 1 by increments of 0.05. A simple representation of the steps in the PLSR computation is presented as a flowchart in Fig. 2.9.

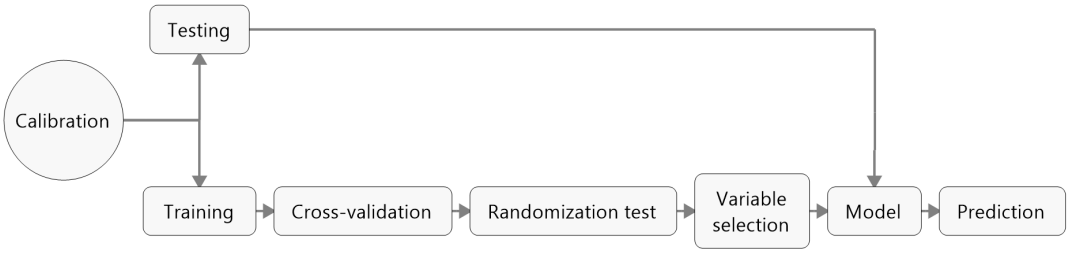


Figure 2.9: Flowchart showing the general structure of the PLSR calculation.

Based on the result from the PLSR and the observations made in PCA, two different approaches were taken to improve the performance of the PLSR model. A PLSR computation without the eleven points seen as separated from the rest of the data in PCA was performed. And a computation where only object with experimental ΔE_{st} values measured in toluene was performed. Note that the percentage of objects reserved as test set was changed to 20 % and 15 %, respectively. For both these computations all the descriptors were included.

Three Cubist regression tree models were created using an in-house script. The script utilizes the R packages `Cubist` [90], `caret`[91] and `doParallel`[92]. The input files are the preprocessed descriptors and the experimental ΔE_{st} values. Note that, contrary to the PLSR calculations, the predictors were not altered. Only the REST set of descriptors were considered. In addition to computation with the full data set, separate computations where done with the eleven objects excluded and only toluene as solvent included. Prediction were done using the same objects which were used to build the model. Then the Spearman rank correlation were calculated.

2.4 Density functional theory

Because of time constraint only a few structures were chosen for further study using DFT and TDDFT. The selection was based on chemistry, size and ΔE_{st} values. A histogram showing the distribution of the ΔE_{st} values in the data set is given in Fig. 2.10. K-means cluster analysis, described in Section 2.2, was used to separate the structures based on chemistry. The size in terms of number of atoms was also an important factor given that calculations on large molecules are more computationally demanding. These three tings taken into account, the structures given in Table 2.2 were selected.

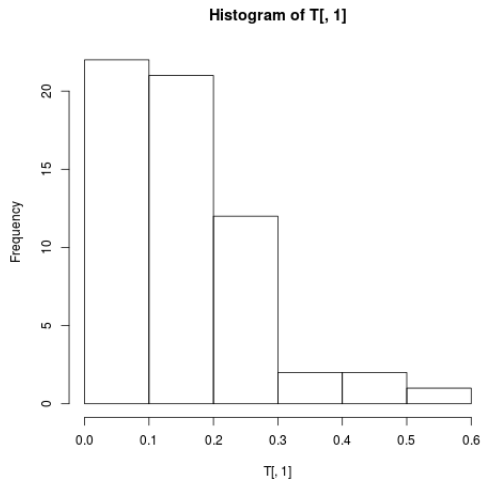


Figure 2.10: Distribution of ΔE_{st} values in the data set. The x-axis represents the intervals the ΔE_{st} lies within. The y-axis shows the frequency of ΔE_{st} values which lie within each interval.

Table 2.2: Selected structures for further DFT calculations based on distribution of ΔE_{st} values, size and chemistry.

Structure name	ΔE_{st}	Cluster affiliation	No. atoms
1	0.54	C1	69
3	0.32	C2	113
2DAC-Mes3B	0.058	C8	122
2PXZ-TAZ	0.23	C3	80
BCzT	0.31	C3	89
BTCz-2CN	0.17	C5	72
Cz2BP	0.21	C4	64
SPXZPO	0.26	C6	56
TCzTrz	0.16	C9	99
TMCPOB	0.09	C7	92

The DFT and TDDFT were performed on a computer with Intel Xeon E5-2687W v2, 3.4 GHz CPU and 198 GB RAM and computations were done using the chemical computation program NWChem [93]. A geometry optimization of each structure in gas phase were performed using DFT with the s12g functional and pcs-0 basis set. The frequencies were also calculated in order to ensure that ground state was achieved. The output coordinates where then used in geometry optimization and frequency calculations in the COSMO solvation model [94, 95]. The s12g functional and pcs-1 basis set was used. The output coordinates were then used in TDDFT with the COSMO solvation model, CAM-B3LYP functional and aug-pcs-1 basis set. Unfortunately, there were some issues running the last calculation and due to time constraints these calculations were not completed.

3 Methods

3.1 Principal component analysis

Principal component analysis is a data compression method where the goal is to express the main information in the observed variables $\mathbf{X}$ in a few latent variables called scores or principal components (PC), $\mathbf{T}$ [96]. Compression of the data is done by projecting the observed data onto a lower dimensional space by means of the projection matrix $\mathbf{P}$, also called the loading matrix [97, 98]. The scores matrix comprises the new coordinates for the observed data and the loading comprise the direction coefficients of the PCs. PCs are constructed such that the first PC points in the direction of maximum variance. The second PC is orthogonal to the first PC and points in the direction of maximum variance not explained by the first PC, and so forth. The mathematical least squares model relation between the observed data $\mathbf{X}$, the scores $\mathbf{T}$ and the loadings $\mathbf{P}$ is defined as

$$\mathbf{X} = \mathbf{T}_A\mathbf{P}'_A + \mathbf{E}_A \tag{3.1}$$

where the subscript A indicates that matrices has been generated using A principal components [97, 98]. The E_A matrix is the residual matrix which contains information about the variance not explained by the model part $\mathbf{T}_A\mathbf{P}'_A$. The model shows that the scores can be viewed as linear combinations of the variables with coefficients $\mathbf{p}'_a$ and conversely, the loadings as a linear combinations with coefficients $\mathbf{t}_a$. Thus, PCA is categorized as a bilinear model (BLM). Before a PCA is performed the observed data is usually centered and scaled in order to let each variable contribute equally to the PCA model.

Compression of large amounts of observed data by utilizing the PCA method increases the interpretation and visualization possibilities. By plotting the scores, e.g. $\mathbf{T}_2$ against $\mathbf{T}_1$, clusters, patterns and possible outliers can be detected. Variables with large score along a PC is important for its direction. A loadings plot can give information about the variable contribution to the model and variable correlation. Unimportant variables clusters at the origin. The direction of the

loadings and the angle between them from the origin determines their correlation. An angle of 0° indicate positively correlated variables, 180° negatively correlated and 90° non-correlated.

3.2 Partial least squares regression

Partial Least Squares Regression (PLSR) is a bilinear modelling method for multivariate calibration [97, 96]. It can yield reliable predictors $\widehat{\mathbf{Y}}$ by projecting many variables $\mathbf{X}$ onto few latent variables $\mathbf{T}$ and then using $\mathbf{T}$ as regressors for $\mathbf{Y}$. The hat notation denotes predicted values as opposed to observed values. The $\mathbf{X}$ matrix contains the independent variables and $\mathbf{Y}$ contains the dependent variables. It is usually assumed that there is a causal relation between $\mathbf{X}$ and $\mathbf{Y}$. Thus, $\mathbf{Y}$ is actively used in the decomposition of $\mathbf{X}$ such that the latent variable projection of $\mathbf{X}$ is directly relevant for the prediction of $\mathbf{Y}$. The matrix structure of the bilinear structure model of $\mathbf{X}$ and $\mathbf{Y}$ is depicted in Fig. 3.1 [99]. Note that $\mathbf{Y}$ can be expressed directly from $\mathbf{X}$ by the regression coefficient $\mathbf{B}$.

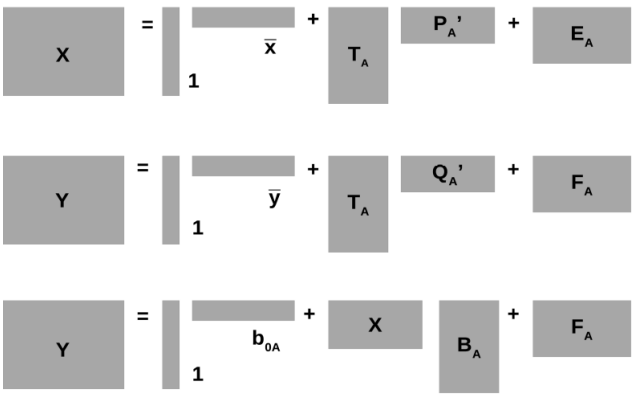


Figure 3.1: Bi-linear structure model for $\mathbf{X}$ and $\mathbf{Y}$

The observed X- and Y-variables are usually centered and scaled to unit variance, corresponding to giving each variable equal weight[100, 101, 97, 96, 102, 99].

3.2.1 Validation

For calibration methods, validation is of vital importance. The validation of the model is done to assess the models predictive ability and to determine the optimal

number of components [101, 96]. If too few components are included in the model it means that important information in the data is not captured by the model, resulting in an underfitting. If too many components are included, which increases the model complexity, it results in an overfitting and noise is affecting the model [97]. The determination of optimal number of components is a balance between reducing the complexity of the model and reducing the error. If the choice is between adding one more component for a minimal reduction of error and not including the component, the component is most likely not included because this will increase the dimensions and thus the complexity of the model. One way of validating the model is by the method of n-fold internal cross-validation. The method consists of dividing the objects in the calibration data into n segments. One segment is left out, retained as a test set. A model is then created from the remaining training set with the one segment left out. The process is repeated until all segments have served as a test set. The n models produced can then be combined to produce a single model. Note that when a model is constructed for a cross-validation segment with few objects it will often perform worse than the model with all the objects included in the calibration data [97].

3.2.2 Randomization testing

The statistical method of randomization testing, also called permutation testing, is based on destroying an assumed relationship between the $\mathbf{X}$ and $\mathbf{Y}$ data by shuffling the dependent variables [97]. By doing so, systematic error can be eliminated [103]. For each principal component in PLSR a test statistic is calculated e.g. the covariance between the scores and the y-values[104]. By permuting the y-values a distribution of test statistics is obtained. Then a parameter α is calculated as the number of times the test statistics of the permuted y-values is equal to or higher than the statistic of the unperturbed situation. If α is larger than a predetermined significant level the component is seen as not being significantly different from noise and should not be included in the model.

3.2.3 Performance metrics

There are several different statistics used to describe the performance of a calibration model. Here, the variable importance for projection (VIP), mean absolute error (MAE), correlation coefficient R^2 and root mean square error of prediction (RMSEP), but there are many others that can be used and evaluated. The VIP parameter is frequently used to evaluate how important a variable is for the modelling of both $\mathbf{Y}$ and $\mathbf{X}$. VIP is a weighted sum of squares of the PLS-weights, where the weights are calculated from the amount of Y-variance of each PLS component[101, 105]. The VIP value is calculated for each variable, thus its can be used as a means for variable selection. As a rule of thumb, a VIP value below 1 is considered as an unimportant variable. However, the model appearance must be checked before blindly leaving out all variables below 1 [105]. Note that VIP is restricted to PLSR and can not be used in e.g. Cubist.

MAE is defined as

$$MAE = \frac{1}{I}\sum_{i=1}^{I}|y_{i,obs} - y_{i,pred}| \tag{3.2}$$

where $y_{i,obs}$ is the observed y-value for object i, $y_{i,pred}$ is the predicted y-value for object i and I is the total number of objects. The MAE is sometimes preferred as a measure of average error compared to RMSEP because it does not deal with squares. Large errors have a greater impact on the total sum of squares, making statistics such as RMSEP harder to interpret [106]. The RMSEP is expressed as

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I}(y_{i,obs} - y_{i,pred})^2}{I}} \tag{3.3}$$

The correlation coefficient is a measure of the predictive ability of a model and is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{I}(y_{i,obs} - y_{i,pred})^2}{\sum_{i=1}^{I}(y_{i,obs} - \bar{y}_{obs})^2} \tag{3.4}$$

where $\bar{y}_{obs}$ is the mean of the observed y-values [69, 107]. When the model is calibrated by cross-validation the correlation coefficient is denoted R^2_{CV} and $y_{i,pred}$ is replaced by cross-validated prediction of the y-variables. R^2 and R^2_{CV} have different value ranges, $0 < R^2 < 1$ and $-\infty < R^2_{CV} < 1$ [107]. A high R^2 value

indicates a good predictive model.

A standard method for outlier detection in PLSR is plotting the studentizes residual against the leverage. The leverage considers the position of the object's X-variables relative to each other. It is proportional to the Hotellings T^2 statistic and the object's Mahalanobis distance measured from the centroid of the training set [96, 97, 108, 109]. It is defined as

$$H_A = X_A(X'_A X_A)^{-1} X'_A \tag{3.5}$$

Here the subscript A is included to emphasis that the leverage is dependent on the number of principal components used. The diagonal element h_{ii} of matrix H_A is the leverage of object i in the data set. A threshold is set as the distance corresponding to a 95th percentile. Objects with a leverage higher than the threshold value are considered unreliable and may be potential outliers. Studentized residual of object i is defined as

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}} \tag{3.6}$$

where e_i is the y-residual and here MSE is the mean square error of the y-residuals [110, 96].

Sprearman's rank correlation coefficient, r_s, is a statistical measure of the strength of a monotonic relationship between variables [111, 112]. Given two samples X and Y of size n the Spearman coefficient can be defined as

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} (R_{X_i} - R_{Y_i})}{n(n^2 - 1)} \tag{3.7}$$

where R_{X_i} and R_{Y_i} is the rank of object i compared to the other values in the sample X and Y respectively. If X_i is the smallest value in the sample, then it would have rank equal to 1. r_s takes values between -1 and 1 where $r_s = 1$ represents perfectly positive correlated samples.

3.3 Cubist

Cubist is a rule-based model that merges several methodologies introduced by Quinlan [113, 114, 115]. The method distinguish itself from other decision trees, such as random forrest and bagging trees, in the techniques used for linear model smoothing, creating rules and pruning[116]. Also, the model consists of an optional boosting-like procedure called committees. The tree is constructed by splitting the data set based on the expected reduction in the node's error rate. The split that is associated with the largest reduction in error is chosen. Then a linear model is created at each node using the splitting variable at that node and all the preceding splitting variables at parenting nodes. The tree growing process continues along the branches until there are no further improvements in error rate or there are not enough samples to continue. When the complete set of linear models have been generated the tree is simplified. For each linear model at the nodes, an adjusted error rate is computed. The adjusted error rate is calculated by removing one model term at a time. Only the model terms that result in a decrease in error rate are removed from the linear model. After the simplification process, yielding the final linear model at each node, the tree undergoes a smoothing process. The models in each branch are combined using a linear combination of two models

$$\hat{y}_{par} = a\hat{y}_k + (1-a)\hat{y}_p \tag{3.8}$$

where a is the smoothing coefficient, $\hat{y}_k$ is the prediction from the current model and $\hat{y}_p$ is the prediction from the parent node. The smoothing coefficient is expressed as

$$a = \frac{Var(e_p) - Cov(e_k, e_p)}{Var(e_p - e_k)} \tag{3.9}$$

where e_k is the residuals from the child model and e_p is the residuals for the parent model. The model with the smallest RMSE has a higher weight in the smoothed model. The final model tree is used to construct the initial set of rules where one linear model is associated with each rule.

In addition to the composite rule-based model, committees can be created which are made up of several rule-based models. The models in the committee are affected

by the result of the previous model. The first member of a committee is always the same as the rule-based model without the committee. The second committee member is a rule-based model which is aimed at improving the prediction of the first member, and so forth. This means that the training set outcome is adjusted based on the prior model fit and then builds a new set of rules using this pseudo-response. The adjusted response for the mth committee model can be expressed as

$$y_m^* = y - (\hat{y}_{m-1} - y) \tag{3.10}$$

where $\hat{y}_{m-1}$ is the predicted value of the prior model, y is the observed value and y_m^* is the adjusted response. Each member of the committee predicts the target value and then the committees' predictions are averaged to give a final prediction. Committee models are in general more suited for fine-tuning a good model rather than overcoming the deficiencies of a poor model.

Once the rule-based model is finalized, Cubist can further adjust the model prediction with samples from the training set. When a new sample is predicted, the K most similar neighbours are determined from the training set. The prediction of a new sample can then be determined based on the neighbours' observed and predicted value and a calculated weight factor.

$$\text{Final prediction} = \frac{1}{K}\sum_{t=1}^{K} w_l[t_l + (\hat{y} - t_l)] \tag{3.11}$$

t_l is the observed value for a training set neighbour, $\hat{t}_l$ is the predicted value of the training set neighbour, $\hat{y}$ is the predicted value of the new sample and w_l is a weight factor. The weight factor is calculated using the distance between the training set neighbour, D_l, and the new sample.

$$w_l = \frac{1}{D_l + 0,5} \tag{3.12}$$

The distance used in these calculations are Manhattan distances. Also, in order to filter the neighbors, the average pairwise distance of the data points in the training set is used as a threshold.

3.4 K-means cluster analysis

Cluster analysis has the goal of separating a collection of objects into clusters such that objects within a cluster are more closely related to one another than objects in other clusters[97, 117]. Central to cluster analysis is the concept of similarity. Similarity is a measure of how close two instances are to each other. In K-means clustering the similarity measure is the Euclidean distance

$$d(x_i, x_{i'}) = ||x_i - x_{i'}||^2 \tag{3.13}$$

The K-means algorithm requires an estimate of the N number of clusters before finding the centre points of these clusters. Each object is assigned to the cluster it is closest to based on the distance between that object and the cluster centre. For each cluster the mean distance between the objects and the centre is computed which is ascribed as the new cluster centre. The cluster assignment and mean distance is itteratively calculated until converged. The main problem with this method is that the optimal number of clusters are not known*a priori* and has to be guessed. A method based on gap statistics enables the determination of the optimal number of clusters. The optimal number of clusters is determined such that the within cluster sum of squares is minimized. Generally the within cluster sum of squares decreases with increasing N.

3.5 Density functional theory

The density functional theory (DFT) is a computational procedure for molecular electron structure calculations[118, 17]. Unlike wave function methods which use the wave function to represent the electronic system, DFT uses the electron density. The energy of the system can be uniquely determined as a functional of the electron density. One advantage using DFT instead of wave function approaches is that the number of variables are independent of the size of the system whereas for wave function methods the complexity increases exponentially. Perhaps the biggest problem with DFT is the functional that connects the density and the ground state energy. The functional in question is the exchange-correlation func-

tional, which is unknown and needs to be approximated. The exchange-correlation functional will be discussed later in this section along with the fundamental theorems of Hohenberg-Kohn, Kohn-Sham equation, approximations of the exchange-correlation functionals and basis sets.

The proof that the electron density determines the properties of a molecule was given by Hohenberg and Kohn in 1964[119]. Their theorems represents the foundation of all modern density functional theories. The first Hohenberg-Kohn (HK) theorem, often referred to as the Hohenberg-Kohn existence theorem, states that the ground state energy and all other ground state electronic properties are uniquely determined by the electron density[17]. Quoting the Hohenberg-Kohn paper:"the external potential $V_{ext}(\vec{r})$ is (to within a constant) a unique functional of $\rho(\vec{r})$; since, in turn $V_{ext}(\vec{r})$ fixes $\hat{H}$ we see that the full many particle ground state is a unique functional of $\rho(\vec{r})$"[119]. The second Hohenberg-Kohn theorem, often called the Hohenberg-Kohn variational theorem, states that the energy functional cannot be less than the true ground state energy of the molecule, and delivers the ground state energy if and only if the density is the true ground state density [17, 120]. This is the variational principle; the obtained energy is an upper bound to the true ground state energy. In practice the variational principle does not hold as an approximated universal functional is used, which by extension means that the Hamiltonian is approximated. The HK theorems provide no means of accessing the density that determines the ground state energy, only that the connection exists.

In 1965 Kohn and Sham derived a set of equations which enabled the electron density to be determined [121]. The derivation was done by considering a hypothetical reference system consisting of N non-interacting electrons in an external potential selected so that the electron density of the reference system was the same as the true electron density[120]. This made it possible to calculate the major part of the kinetic energy to a good accuracy and then approach the remaining part in an approximate manner. The non-interacting kinetic energy is not equal to the true kinetic energy of the interacting system, even if the systems share the same

density. Kohn and Sham accounted for that by separating the universal functional

$$F[\rho(\vec{r})] = T_S[\rho(\vec{r})] + J[\rho(\vec{r})] + E_{xc}[\rho(\vec{r})] \tag{3.14}$$

where $E_{xc}[\rho]$ is the exchange-correlation energy containing everything that is unknown, T_S is the non-interacting kinetic energy and J is the Coulomb interaction. The exchange-correlation energy is here defined as

$$E_{xc}[\rho] = T_C[\rho] + E_{ncl}[\rho] \tag{3.15}$$

where T_C is the true kinetic energy and E_{ncl} is a non-classical contribution. Considering the energy of the real interacting system the one-electron Kohn-Sham (KS) equation can be expressed as

$$\hat{f}^{KS}\phi_i = \varepsilon_i\phi_i \tag{3.16}$$

with the one-electron Kohn-Sham operator $\hat{f}^{KS}$ being defined as

$$\hat{f}^{KS} = -\frac{1}{2}\nabla^2 + V_{eff}(\vec{r}) \tag{3.17}$$

V_{eff} is the effective potential including the exchange-correlating potential, which is defined as the functional derivative of the exchange-correlation energy with respect to the density.

As mentioned earlier, all the unknown contributions to the electronic energy are collectively folded into the exchange-correlation functional. These contributions cannot be determined exactly in the DFT environment, thus approximations are necessary. The quality of the density functional approach is solely dependent on the accuracy of the chosen approximation to E_{xc}. Numerous schemes have been developed in order to approximate the exchange-correlation functional and the search for better and improved approximations is at the very center of DFT research. Here, the local density and the local spin-density approximations will be considered along with the generalized gradient approximations and hybrid functionals.

Virtually all approximate exchange correlation functionals are based on the idea of a hypothetical uniform electron gas. This is a system where electrons move on a positive background charge distribution such that the total ensemble is electrically neutral. This type of system is far from the real situation where densities rapidly change within the molecule. The reason for using it in DFT is that it is the only system where the form of the exchange and the correlation energy functionals are known exactly or at least to a very high accuracy.

The uniform electron gas system is utilized in the local density approximation (LDA). Central to the LDA is the assumption that the exchange-correlation functional can be split into an exchange functional and a correlation functional. An explicit expression can be found for the exchange functional, but not for the correlation functional. Analytical expressions for the correlation functional have been proposed by many different authors. The local spin-density approximation (LSD) is an extension of the LDA to an unrestricted case. Here the local spin-densities, $\rho_\alpha(\vec{r})$ and $\rho_\beta(\vec{r})$, are used as primary input instead of the electron density. The benefit of using the LSD is increased flexibility as a result of having two variables instead of just one. Also, for open shell systems with even or uneven number of α and β electrons, LSD can perform better due to symmetry breaking.

Further extension of local approximations propose that the exchange-correlation functional should not only depend upon the electron density at a certain position, but also the gradient. The inclusion of the gradient account for the non-homogeneity of the true electron density. Functionals that include the gradient of the electron density are collectively known as generalized gradient approximations (GGA) and are the workhorses of DFT. These functionals also include some restrictions regarding the exchange and correlation holes, which separates these functionals from the gradient extension approximation. The GGA exchange-correlation energy functional is split into exchange and correlation contributions and approximations for each are sought separately. A popular exchange functional, referred to as B or B88, was developed by Becke in 1988[122]. It was designed to recover the exchange energy density asymptotically far from a finite system. Other functional based on the same principles are the CAM(A) and CAM(B) function-

als developed by Handy, Thermath and Laming in 1993 [123]. A corresponding correlation functional which is very much used today is the LYP functional by Lee, Yang and Parr [124]. This functional distinguishes itself from the rest of the functionals mentioned. It is not based on the uniform electron gas but on wave function based theory by Colle and Salvetti [125]. It should also be noted that this correlation functional only includes dynamical correlations effects [120].

In principle, each exchange functional could be combined with any of the correlation functionals. However, only a few combinations are used. The exchange functional is almost always chosen to be Becke's (B) and a popular combination is with the LYP correlation functional. This combination is referred to as the BLYP functional. Usually the exchange contributions are significantly larger in absolute numbers than the corresponding correlation effects. In order to get good results from the density functional theory, it is important to express the exchange functional accurately.

Hybrid functionals are functionals of which the exchange-correlation energy is expressed as a linear combination of multiple exchange and correlation contributions. It is a hybrid because it includes pure density functionals and exact Hartree-Fock exchange. Currently, the most popular hybrid functional is known as B3LYP and was suggested by Stephens et al. in 1994 [126]. The B3LYP exchange-correlation energy expression is

$$E_{xc}^{B3LYP} = (1-a)E_X^{LSD} + aE_x^{HF} + bE_X^{B88} + cE_C^{LYP} + (1-c)E_C^{LSD} \tag{3.18}$$

where a, b, and c are semi-empirical coefficients to determine the weight of each contribution. a determines the amount of exact exchange, b and c determine the exchange and the correlation gradient corrections to the local density approximation. E_x^{HF} is the exact Hartree-Fock exchange energy, E_x^{LSD} and E_c^{LSD} are the local exchange and correlation functionals from the LDS approximation, respectively. E_X^{B88} is Becke's gradient correction to the exchange energy and E_c^{LYP} is Lee, Yang and Parr's gradient correction to the correlation functional. Although it performs better than LDA and other combinations of functionals, B3LYP has some short

comings. In practical applications it is unsuccessful in describing the polarization of long chains, excitations using TDDFT for Rydberg states [127, 128, 129] and charge transfer excitations[130, 131, 132]. The reason for these failings are the behavior of the exchange potential at long range. Yanai, Tew and Handy proposed using the range-separated Coulomb-attenuating method (CAM)[133] in combination with the hybrid functional, i.e. CAM-B3LYP, to overcome these short comings. Another type of GGA functional is the s12g functional developed by Marcel Swart [134]. It includes a dispersion contribution in order to improve the description of weakly bound systems without any computational cost.

3.5.1 Basic machinery of density functional theory

Some of the most central strategies for making the Kohn-Sham scheme computationally manageable are presented here. At the core of DFT is the one-electron Kohn-Sham equation as given in Eq. (3.16). This equation is handled by applying the linear combination of atomic orbitals (LCAO) to a finite set of basis functions in order to expand the KS-orbitals. The full one-electron KS-equation can be expressed as

$$\left(-\frac{1}{2}\nabla^2 + \left[\sum_j^N \int \frac{|\varphi_j(\overrightarrow{r_2})|^2}{r_{12}} d\overrightarrow{r_2} + V_{xc}(\overrightarrow{r_1}) - \sum_A^M \frac{Z_A}{r_{1A}}\right]\right)\varphi_i = \varepsilon_i \varphi_i \tag{3.19}$$

The KS-orbitals is expressed as a linear combination of the predefined basis functions η_μ

$$\varphi_i = \sum_{\mu=1}^{L} c_{\mu i}\eta_\mu \tag{3.20}$$

Then the KS-equation can be written in matrix notation as

$$\mathbf{F^{KS}C = SC\varepsilon} \tag{3.21}$$

where $\mathbf{F^{KS}}$ is the KS-matrix, $\mathbf{S}$ is the overlap matrix, ε contains the orbital energies and $\mathbf{C}$ contains the expansion coefficients. Through the LCAO expansion the problem has been translated from a non-linear one to a liner one which can easily be implemented in computer programs. By expanding the Kohn-Sham operator

into its components, the individual elements of the Kohn-Sham matrix become

$$F_{\mu\nu}^{KS} = h_{\mu\nu} + J_{\mu\nu} + V_{\mu\nu}^{xc} \tag{3.22}$$

where $h_{\mu\nu}$ is the one-electron contribution which describes the electronic kinetic energy and the electron-nuclear interaction. $J_{\mu\nu}$ is the Coulomb contribution while $V_{\mu\nu}^{xc}$ represents the exchange-correlation contribution. A simplified expression for the Coulomb contribution is

$$J_{\mu\nu} = \int\int \eta_\mu(\vec{r_1})\eta_\nu(\vec{r_1})\frac{\rho(\vec{r_2})}{r_{12}}d\vec{r_1}d\vec{r_2} \tag{3.23}$$

The exchange-correlation part of Eq. (3.22) can be expressed as

$$V_{\mu\nu}^{xc} = \int \eta_\mu(\vec{r_1})V_{xc}(\vec{r_1})\eta_\nu(\vec{r_1})d\vec{r_1} \tag{3.24}$$

An analytical expression for V_{xc}, even for the simplest approximations such as LDA, is out of reach due to the complicated mathematics. Numerical techniques based on a grid is therefore employed to solve these integrals. Once the suited grid is chosen, the exchange-correlation potential must be evaluated at each grid point. Most computer programs follow the design principles of Becke [135], where the integrals are broken up into separate but overlapping atomic contributions. Once the atomic contributions are determined, the corresponding integrals are computed on grids which comprise of points on concentric spheres around each atom. By converting to polar coordinates the integration can be separated into an angular contribution and a radial contribution. There are many different numerical quadratures available for both the angular and the radial contributions, and different computer programs utilize different schemes. E.g. the quantum chemical software NWChem uses the Euler-McLaurin scheme proposed by Murray, Handy and Laming in 1993 [136], with a modified Mura-Knowles transformation [137], as numerical quadrature for integration of the radial part. The numerical quadrature for the angular part is almost always explicitly chosen to be the Lebedev grids [138, 139, 140, 141, 142, 143].

Some of the problematic aspects inherent in the numerical quadrature techniques originates from the fact that none of the numerically approximated quantities are exact. A severe problem is that the total energy of a molecule is not rotationally invariant, i.e. different orientations in space yield different energies. A second major problem connected to the use of finite grids for the evaluation of the exchange-correlation energy is associated with the determination of derivatives of the energy, such as the gradient used in geometry optimizations. The numerical quadrature approximation of the exchange-correlations energy leads to a non-zero gradient at the lowest energy configurations while the structure with vanishing gradient is not the one with lowest energy.

3.5.2 Basis functions

In the Kohn-Sham formalism the orbitals play an indirect role and are introduced only as a tool to construct the charge density according to

$$\rho(\vec{r}) = \sum_{i}^{N} |\phi_i(\vec{r})|^2 \tag{3.25}$$

where ϕ_i is given by Eq. (3.20). There are two types of basis functions commonly used in calculations, Slater type orbitals (STO) and Gaussian type orbitals (GTO) [118]. STOs are mainly used for atomic and diatomic systems where high accuracy is required, and in semi-empirical methods where all three- and four-center integrals are ignored. Density functional methods which do not include exact exchange and where the Coulomb energy is calculated by fitting the density into a set of auxiliary functions are also suitable for the use of STOs. The GTOs are inferior to STOs for two reasons, i) GTOs have problem describing the proper behavior near the nucleus because at the nucleus the GTO has a zero slope in contrast to STOs which have a cusp. ii) The tail of the wave function is represented poorly as GTOs fall off too rapidly far from the nucleus. The consequences of these shortcomings of the GTOs are that, in general, more GTOs are required for achieving a given accuracy compared with STOs. However, the number of GTO basis functions required are compensated for by the computational efficiency. Therefore, GTO basis functions are almost always the preferred basis functions in electronic structure

calculations.

The simplest and least accurate basis set is called the minimum basis set and includes only one basis function for each atomic orbital up to and including the valence orbitals[120, 118]. The next level of sophistication is the double-zeta(DZ) basis sets. Here, the set of functions are doubled, i.e. there are two functions for each orbital. This allows for a more exact description of the electron distribution which can be different in different directions. Taking into account the fact that chemical processes occur in the valence space, the doubling of basis functions can be limited to the valence orbitals. This gives rise to the split-valence type basis sets, where the core orbitals are treated in a minimal set. Typical examples are the 3-21G and 6-31G Gaussian basis sets.

In most applications, basis sets are augmented by polarization functions, i.e. functions of higher angular momentum than those occupied in the atom, e.g. p-functions for hydrogen or d-functions for the first-row elements. Polarization functions have by definition more angular nodal planes than the occupied atomic orbitals and thus ensure that the orbitals can distort from their original atomic symmetry and better adapt to the molecular environment. At the HF level, the polarization functions describe charge polarization and at correlation level they describe the electron correlation. In DFT the polarization functions describe both effects[144]. Polarized double-zeta and split-valence basis sets are the mainstay of routine quantum chemical applications since they usually offer a balanced compromise between accuracy and efficiency[120].

Frank Jensen has developed a design principle to convert general contracted basis sets[145] to segmented basis sets[146]. The conversion is done under the requirement that the total energies of the general and the segmented basis sets are identical[147]. Based on the observation made by Davidson[148], that segmentation schemes eliminates the redundancies in the general contraction, more computationally efficient segmented basis sets can be generated with the full accuracy of the general contracted basis sets. The pcs-n basis sets are segmented versions of the polarization consistent pc-n basis sets [147, 144, 149, 150, 151, 152, 153]. Jensen has shown in his research that DFT calculations using these basis sets pro-

duce the lowest basis set error at a given zeta level. Also, they are among the computationally most efficient basis sets.

3.5.3 Excited states - time dependent density functional theory

TDDFT has become a much used method, specially in quantum chemistry and electronic structure theory for many different reasons. Firstly, the TDDFT is a formally exact method according to the theorems of Runge-Gross[154] and van Leeuwen[155]. Secondly, in analogue with the ground state DFT, TDDFT utilizes a non-interacting reference system to obtain the density of the many-body interacting system, which makes it more computationally efficient and easier to implement in computer programs[120, 156]. Time-dependent density functional response theory considers the response of a system, initially in a stationary state, to a infinitesimal perturbation, assuming adiabatic conditions[127]. The density response can be expressed as a response function of a non-interacting Kohn-Sham system and a frequency-dependent exchange-correlation kernel. This method could be used to extract the excitation energies which lie at the poles of the density-density response function [127, 157]. The details of TDDFT will not be discussed here, but a brief presentation of the foundation will be presented below. The reader is referred to literature such as [127, 158] for a more detail description of these methods.

Runge and Gross extended the Hohenberg-Kohn theorems to time-dependent systems, i.e. that observables of a many-body system could be determined from the time-dependent one-body density [159, 158, 127, 160]. They also showed that there must be a one-to-one mapping between the time-dependent density and the time-dependent external potential. In analogue with the ground state DFT the density of a many-body system is obtained as the density of a one-body non-interacting system in a local effective potential. This effective potential is the time-dependent Kohn-Sham potential, which is a function of the exchange-correlation potential, Hartree-Fock potential and an external part. Here the KS potential is assumed to exist, however Runge and Gross do not provide any proof of its actual existence. Thanks to van Leeuwen it was proven that under certain conditions this potential does in fact exist. As for the ground state DFT, the exchange-correlation potential is unknown and must be approximated. However, in TDDFT this is a

much more complicated case as the potential is dependent on the entire history of the density and the initial state of the interacting and non-interacting system [161, 162, 163]. The adiabatic local density approximation, also referred to as the time-dependent local density approximation, is the simplest approximation and has become the workhorse of TDDFT[158, 128, 127]. In this approximation the exchange-correlation potential of TDDFT is approximated by a ground state DFT functional.

4 Results and discussion

4.1 Principal component analysis

As mentioned in the methodology section, the descriptor groups were applied incrementally. This was done in order to see how the increase in variable amount and variable type would affect the variance explained by the number of components. The variance explained by the ten first PCs for each PCA calculation is presented in Table 4.1. As can be seen, the variance explained does not differ much between the different calculations, with the exception of PCA with the descriptor set REST. The 91 % variance explained for the REST set of descriptors is to be expected as the data contained in these descriptors is strongly correlated. A plot of the variance explained by the ten first PCs for the REST set of descriptors is presented in Fig. 4.1. The PCA result for the other combinations of descriptors is presented in Appendix B.1. Because EVA and EEVA are not easily interpreted in the PCA environment, the main PCA result of interest is where these are excluded. Further discussion of the PCA result is based on the REST set of descriptors, if not specified otherwise.

Table 4.1: Variance explained by the first ten PCs for each PCA computation with different combinations of descriptors. Obtained during the master project, fall 2017.

Descriptor	Variance (%)
EVA	50.4
EEVA	47.1
EVA + EEVA	47.1
EVA + REST	51.1
EEVA + REST	48.2
ALL	47.7
REST	91.5

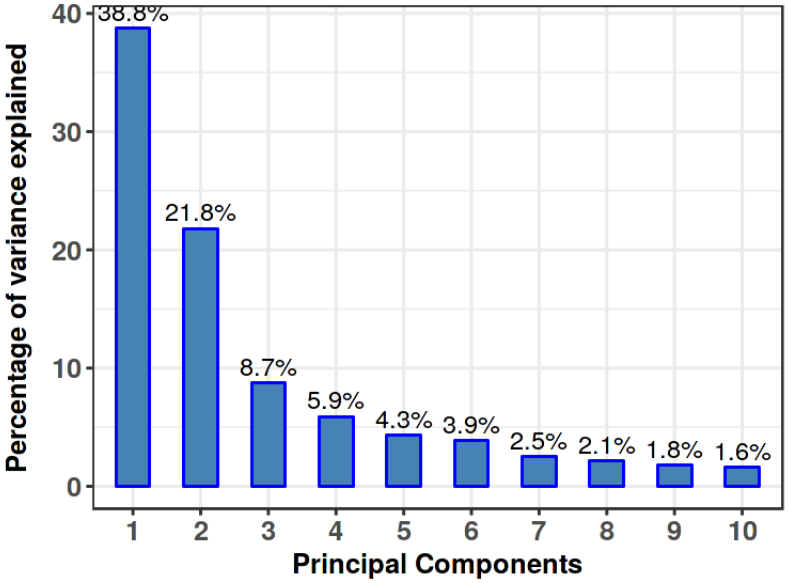
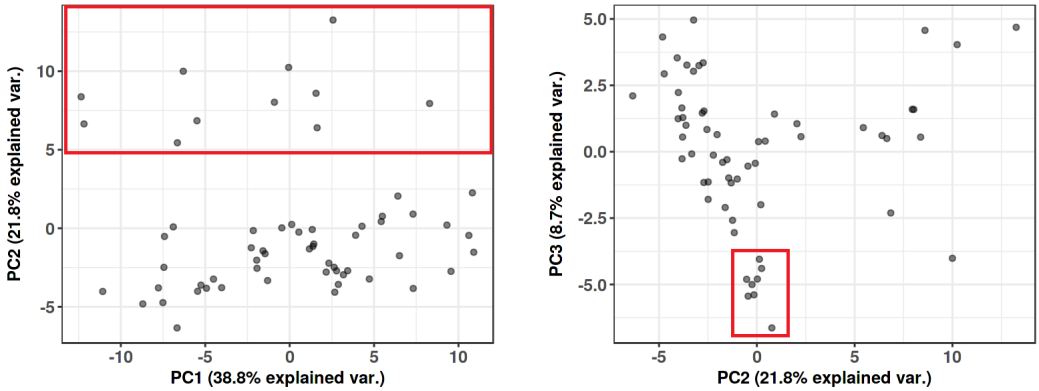


Figure 4.1: Variance explained by the ten first PCs for PCA calculation including the REST set of descriptors. Obtained during the master project, fall 2017.

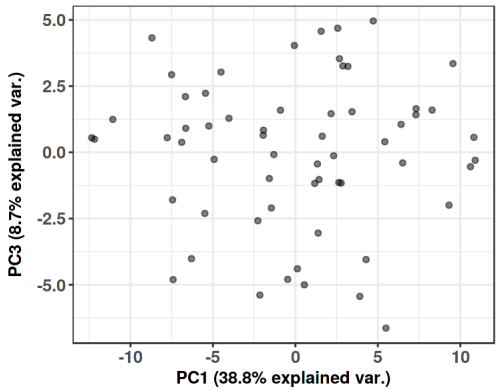
The scores plots for the PCA computation are depicted in Fig. 4.2. The scores plot where PC 2 is plotted against PC 1, see Fig. 4.2a, clearly shows a separation of the data, with eleven points located above a PC 2 value of 5. These eleven points represent the structures 1, 2, 3, DTC-mBPSB, DTC-pBPSB, TPXZPO, DPXZPO, SPXZPO, DMOC-DPS, DMAC-DPS and CzAsSF. In the scores plot where PC 3 is plotted against PC 2, a small cluster of objects can be seen at the bottom of the plot. Including the point at the very bottom, this cluster consists of eight structures, namely DAC-Mes3B, 2DAC-Mes3B, ACRPOB, SXDPAPOB, SFDPAPOB, TMCPOB, TB-1PXZ and TB-2PXZ. In PCA calculations patterns can be seen, but nothing certain can be said regarding clusters and potential outliers. That is one of the reasons the k-means cluster analysis is performed. The result of this analysis will be presented in Section 4.2.

The scores plots reveal an uneven distribution of points which raises some concerns regarding the data set. Fig. 4.2a shows a relatively large empty area, which represents an area where the chemical space is not explained. Such observations are also made in score plot in Fig. 4.2b. The objects and their descriptors are therefore insufficient for describing the full chemical space.



(a) PC 2 plotted against PC 1

(b) PC 3 plotted against PC 2

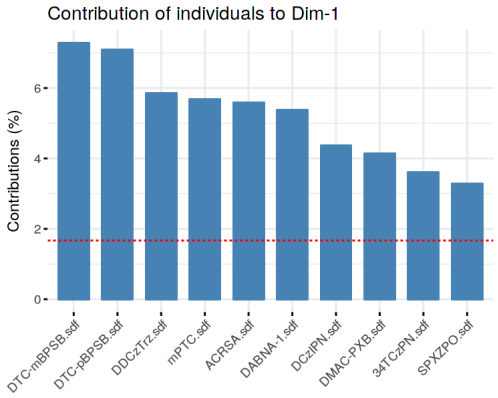


(c) PC 3 plotted against PC 1

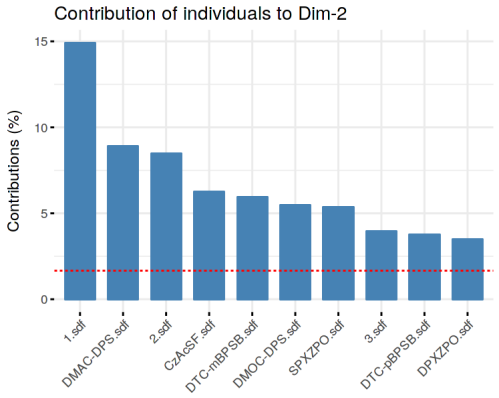
Figure 4.2: Scores plot for when the REST set of descriptors is used and all the objects are included. Obtained during the master project, fall 2017.

Fig. 4.3 is included to illustrate that objects having a high score value along a PC are important for determining the direction of that PC. The ten most contributing objects to PC 2, see Fig. 4.3b, coincides with ten out of eleven points which were

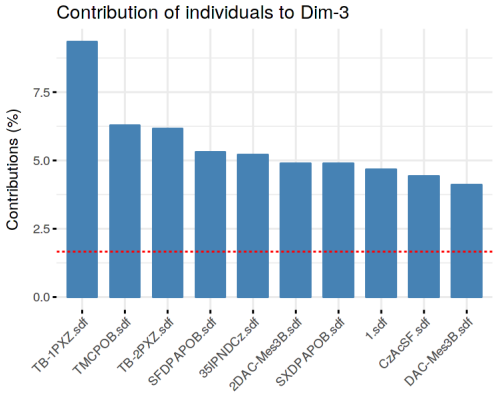
observed separated from the rest of the population. Also, object contribution to PC 3 include seven out of eight points which was observed clustering at the bottom of the scores plot in Fig. 4.2b.



(a) Object contribution to PC 1



(b) Object contribution to PC 2

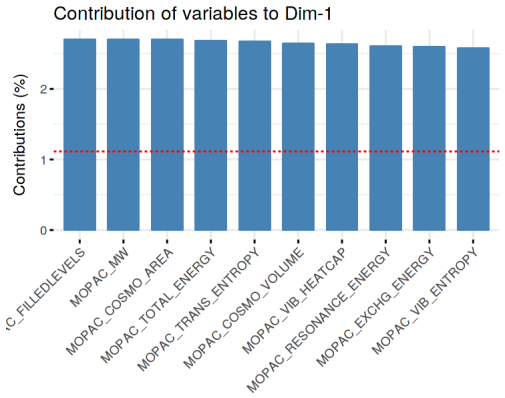


(c) Object contribution to PC 3

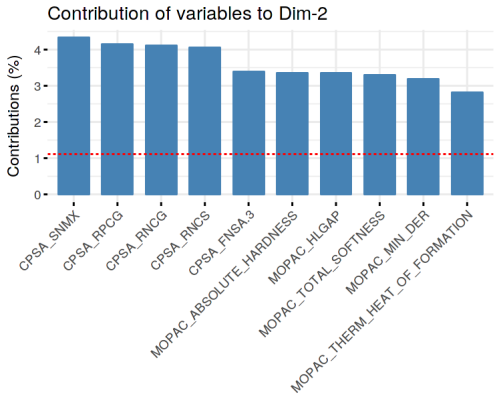
Figure 4.3: Object contribution to the first three PCs when the REST set of descriptors is used and all objects are included. Obtained during the master project, fall 2017.

The ten most contributing variables to the three first PCs are presented in Fig. 4.4. It is interesting to see that for PC 1 only MOPAC descriptors are contributing. For PC 2 and PC 3 a combination of CPSA and MOPAC descriptors are contributing, with the CPSAs being the most dominant ones. Neither geometry or the autocor-

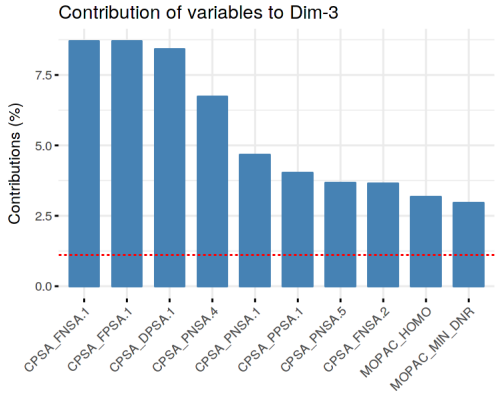
related charge descriptors are included. This does of course not mean that they do not contribute at all, but compared to the MOPAC and CPSA descriptors they contribute only to a fairly small extent.



(a) Variable contribution to PC 1



(b) Variable contribution to PC 2

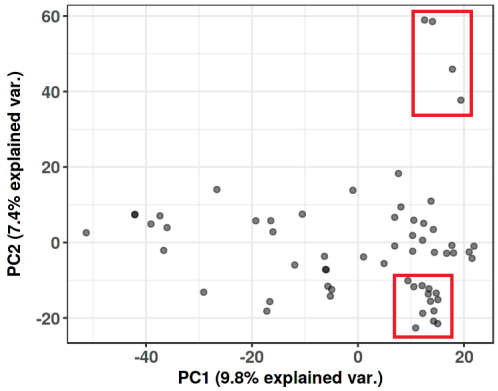


(c) Variable contribution to PC 3

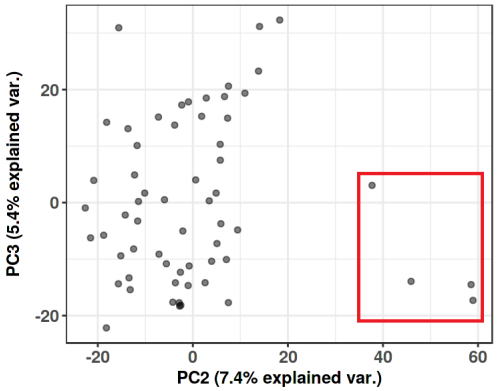
Figure 4.4: Variable contribution to the first three PCs when the REST set of descriptors is used and all objects are included. Obtained during the master project, fall 2017.

The result from the PCA where all the descriptors were included is depicted in Fig. 4.5. There are some points that stand out. E.g. Fig. 4.5a and Fig. 4.5b show four points with large score values for PC 2. The four points represent structure 1, 2, DTC-mBPSB and DTC-pBPSB. By examining the scores plots for EVA,

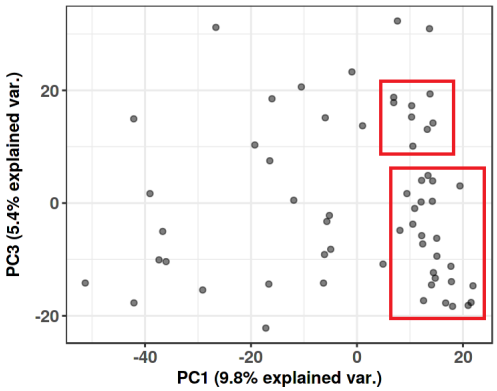
where PC 2 is plotted against PC 1 (see Fig. 4.6), the same four points stand out by having large PC 2 values. As a consequence of EVA and EEVA being spectral like they are not suited for interpretation in a PCA environment, which here is exemplified by how EVA seem to affect the scores in the PCA with all the descriptors included.



(a) PC 2 plotted against PC 1



(b) PC 3 plotted against PC 2



(c) PC 3 plotted against PC 1

Figure 4.5: Scores plot for when all the descriptor and all the objects are included. Obtained during the master project, fall 2017.

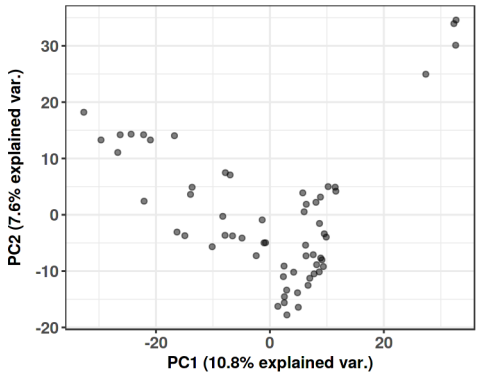
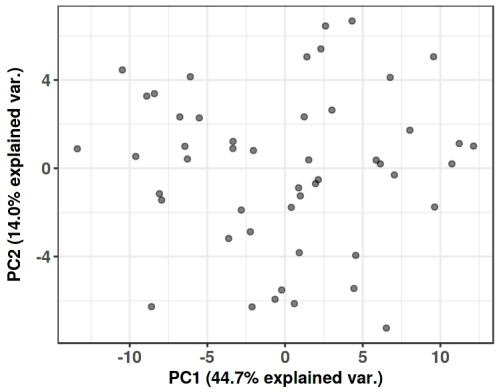
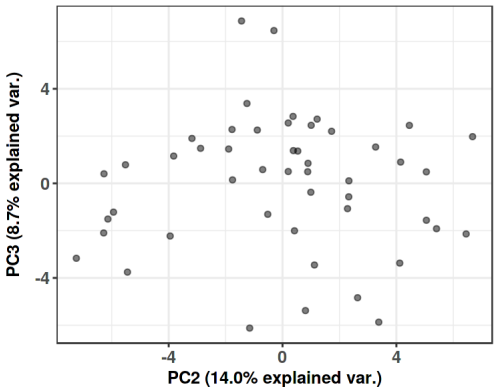


Figure 4.6: Score plot where PC 2 is plotted against PC 1. The EVA descriptors are used and all the objects are included. Obtained during the master project, fall 2017.

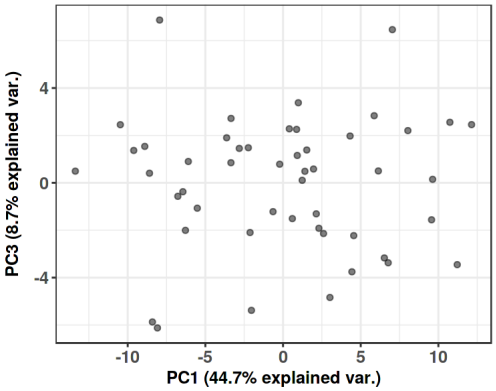
As a result of the uneven distribution of objects and due to poor PLSR results (discussed in Section 4.3), PCA was repeated without the eleven objects separated from the rest of the data. The eleven objects will later be referred to as "the eleven potential outliers". The exclusion of these objects was done especially to see if that would improve the PLSR model. For this case the variance explained by the first ten PCs was 90.3 % and the plot is given in Fig. B.2 in Appendix B.2. The scores plots are presented in Fig. 4.7 with the REST set of descriptors. Here there are two points which stand out, having larger values along the PC 3 axis, see Fig. 4.7b and Fig. 4.7c. These objects represent the Cz2BP and CC2BP structures and can also be seen as the most contributing objects to PC 3 in Fig. B.3iii in Appendix B.2. There is still some empty space associated with the main grouping of objects as can be seen in e.g. the lower left quadrant of Fig. 4.7b. Regarding the most contributing variables, the same trend can be seen as for the original data with the MOPAC and CPSA descriptors being the dominant ones. The variable contributions are given in Fig. B.3 in Appendix B.2.



(a) PC 2 plotted against PC 1



(b) PC 3 plotted against PC 2

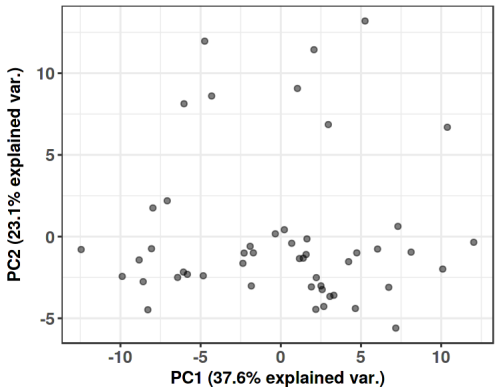


(c) PC 3 plotted against PC 1

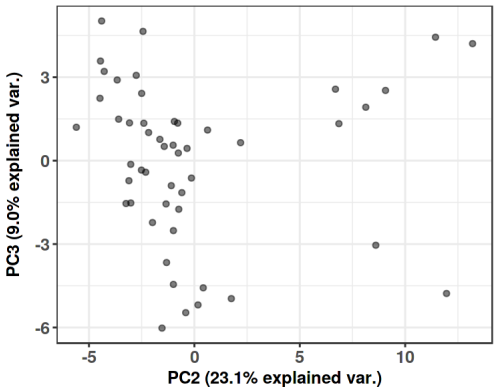
Figure 4.7: Scores plots for when the REST set of descriptors are used and the eleven potential outliers are excluded

Another approach to potentially improve the PLSR model was to only consider structures for which the experimentally determined ΔE_{st} was determined in toluene. The ΔE_{st} values do not affect the PCA, but are important in the PLSR. The PCA is therefore performed as an instructive step in regards to the descriptors. The variance explained by the ten first PCs is 93.8 %, which is a slight improvement from the PCA of the full data set. Notice however that the variance explained by PC 1 is less than that of the full data set. The increase in variance explained is due to an increase along PC 2 and PC 3. The scores plots are given in Fig. 4.8. The

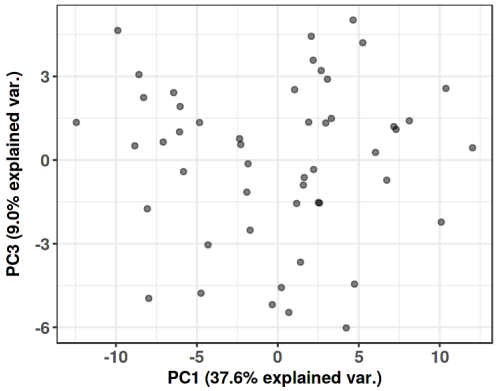
same separation of the data can be seen here as that of the full data set in Fig. 4.2a. One difference is that for the full data set there was eleven points seen as separated from the rest, whereas here there are only eight. Structure DTC-mBPSB, DTC-pBPSB and CzAsSF are among the structures which have been removed. In Fig. 4.2a the two points to the far left in the group of separated points represents DTC-mBPSB and DTC-pBPSB. Removing these may be an explanation for why the variance explained by PC 1 has been reduced in this case. The most contributing variables and objects are presented in Fig. B.7 and Fig. B.6, respectively, in Appendix B.3. In accordance with the PCA result of the full data set, the most contributing objects to PC 2 correspond to the objects separated from the rest of the population. Regarding the variable contributions, the same trend is observed as for the PCA result with the full data set and with the eleven objects removed, only MOPAC and CPSA descriptors are contributing to the first three PCs.



(a) PC 2 plotted against PC 1



(b) PC 3 plotted against PC 2



(c) PC 3 plotted against PC 1

Figure 4.8: Scores plots for when the REST set of descriptors are used and only the structures with experimental ΔE_{st} values measured in toluene are considered

4.2 K-means cluster analysis

All the results from the k-means cluster analysis are presented with the REST set of descriptors. The argument here is the same as for the PCA case, EVA and EEVA are not easily interpreted in this kind of environment. The analysis is done on the full data set as well as without the eleven potential outliers and when only objects with experimental ΔE_{st} values measured in toluene are considered. First considering the result using the full data set, the optimal number of clusters was determined by gap statistics to be 9, see Fig. 4.9. The clusters and cluster affiliations are depicted in Fig. 4.10. Fig. 4.10 is the same as the scores plot in Fig. 4.2a, only with the objects partitioned into clusters. Here, the same partition can be seen as in the PCA result where eleven objects are clearly separated from the rest of the population. The eleven potential outliers are partitioned into three different clusters, namely clusters 3,5 and 9.

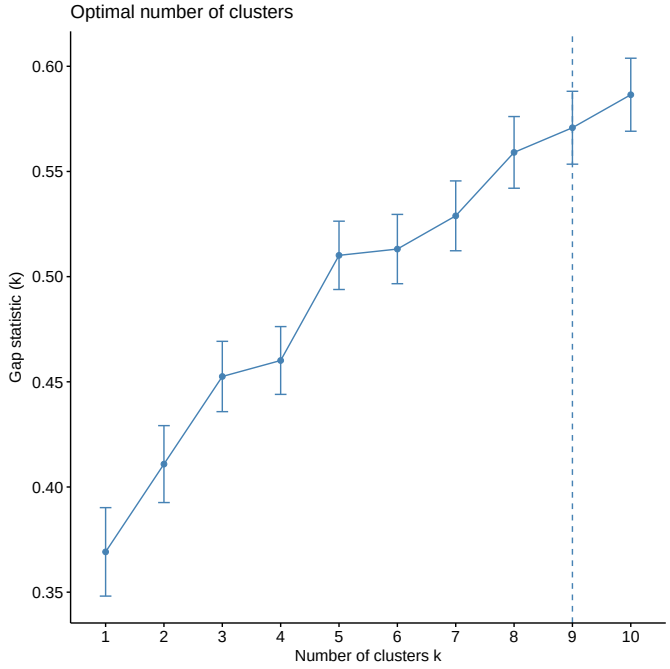


Figure 4.9: Optimal number of clusters when all the structures are included. The y-axis represents the calculated gap statistic for each component. The x-axis represents the number of components.

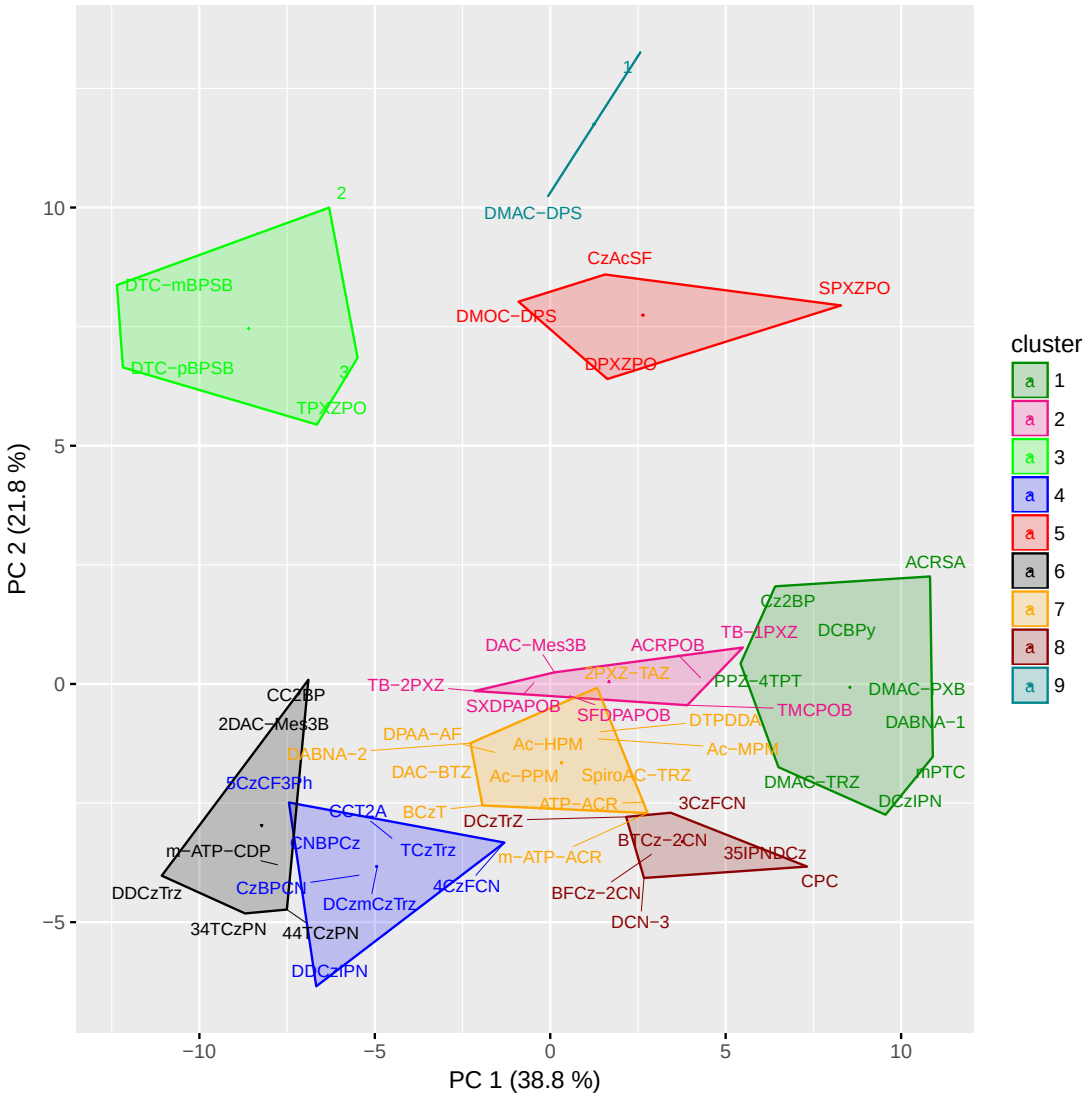


Figure 4.10: The figure display the clusters and the structures affiliation to each of these clusters. The REST set of descriptors are used and all the objects are included. The y- and x-axis represents PC 2 and PC 1, respectively, with the variance explained by each PC given in parenthesis.

What seems to be the main separation factors are the size in terms of number of atoms, shape and charge distribution. The two former factors are evident by inspection of the structures in the different clusters. The latter is more difficult

to see directly from comparing the structures, but it is evident from the main attributing descriptors in the Cubist model discussed in Section 4.4. Also, by inspecting the electrostatic potential surface a clear difference can be observed, especially between the eleven potential outliers and the rest of the population. The members of clusters 3, 5 and 9 all contain sulfonyl groups or phosphine oxide groups which give rise to a charge distribution that is different than for all other structures. An example is DTC-mBPSB in cluster 3, with the highest total dipole moment which is mainly attributed to its two sulfonyl groups. The average total dipole moment and the average charge dipole for clusters 3, 5, and 9 are one average higher than for the other clusters. By inspecting the maximum and minimum values of the descriptors that is shown most attributing in the Cubist model, the CPSA, the total dipole and the charge dipole descriptors are almost exclusively paired with the eleven structures and the rest of the structures. For example, TPXZPO in cluster 3 has the maximum value for the CPSA PPSA-5 descriptor whereas the minimum belongs to DCzIPN in cluster 1.

The fact that there is an optimal number of 9 clusters suggests that the population of structures are highly heterogeneous. All the clusters constitute a variety of different donors and acceptors and have highly different donor-acceptor architectures. This is an undesired situation as the consequence is poor spanning of the chemistry and general variation in the data. 60 objects are considered a small data set, and when these 60 structures display different properties which vary a lot, the data is not sufficient to capture all these variations. The descriptors may also not be sufficient for describing the data and its full variation. It might also be possible that the descriptors chosen in this study as molecular representations of the structures are not the most relevant for predicting ΔE_{st}.

4.2.1 Identification of potential outliers

The optimal number of clusters is again decided by gap statistics to be 5, see Fig. 4.11. Fig. 4.12 displays the clusters along PC 1 and PC 2. Considering only the remaining objects and comparing them with the result where the full data set is used, it can be seen that the number of clusters have been reduced from 6 to 5 and the total reduction is from 9 to 5. This is a positive development, but 5 clusters are still an unsatisfactory number. There are still large regions where no objects can be found and relatively large empty spaces within the clusters as well.

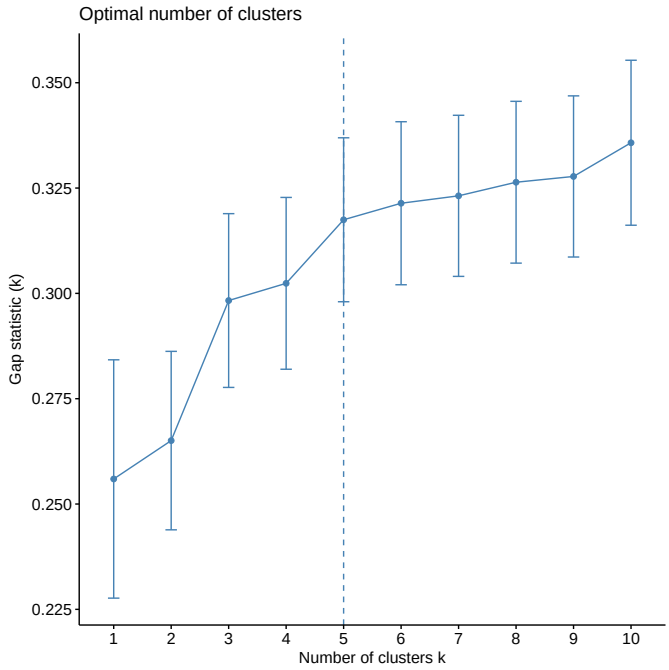


Figure 4.11: Optimal number of clusters when the eleven potential outliers are excluded. The y-axis represents the calculated gap statistic for each component. The x-axis represents the number of components.

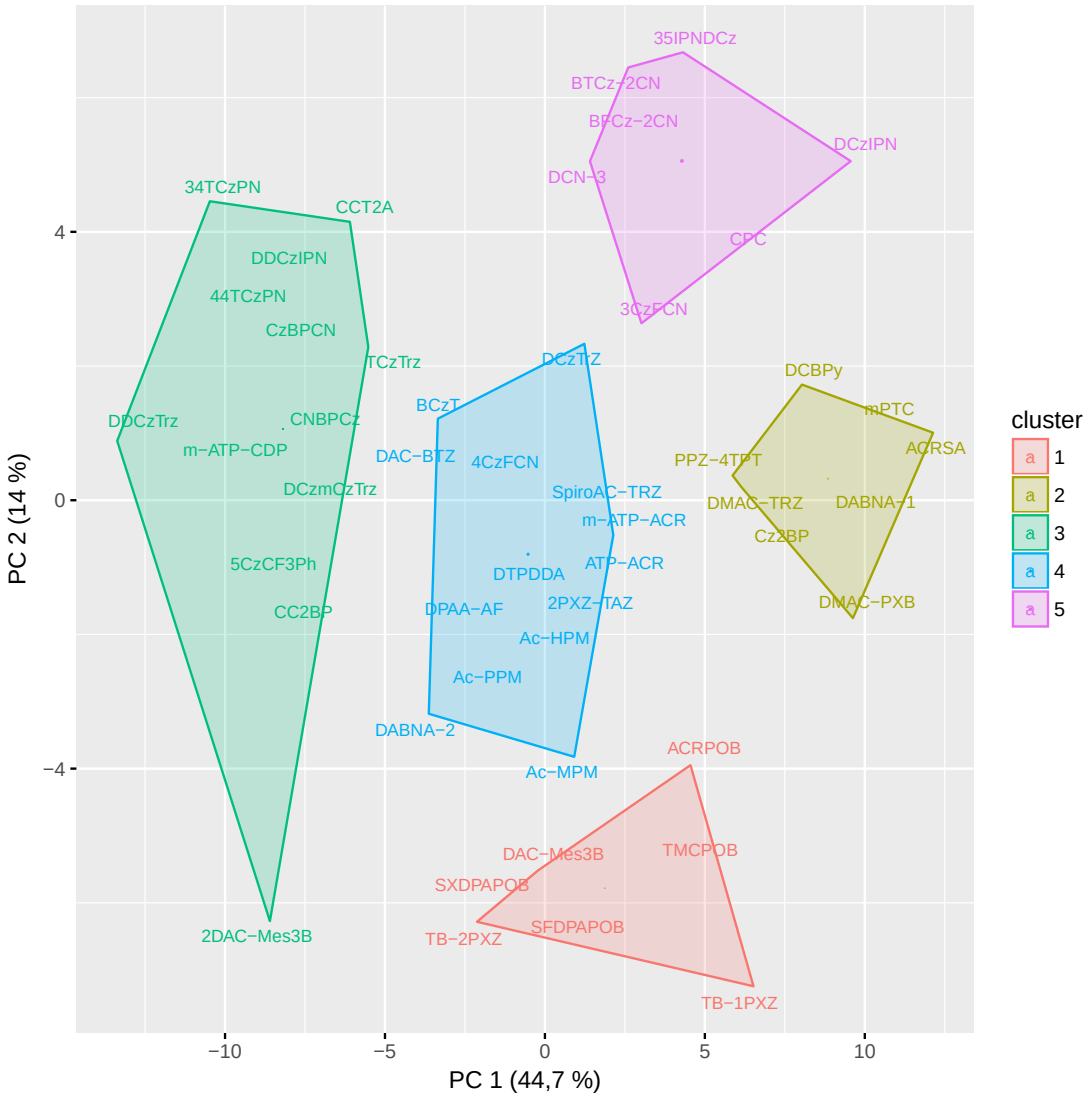


Figure 4.12: The figure display the clusters and the structures affiliation to each of these clusters. The REST set of descriptors are used and the eleven potential outliers are excluded. The y- and x-axis represents PC 2 and PC 1, respectively, with the variance explained by each PC given in parenthesis.

4.2.2 Solvent effects

Optimal number of clusters was here determined to be 9 as can be seen in Fig. 4.13. Cluster affiliations is given in Fig. 4.14. Of course, the same separation of the data can be seen here, although here there are 8 objects instead of 11.

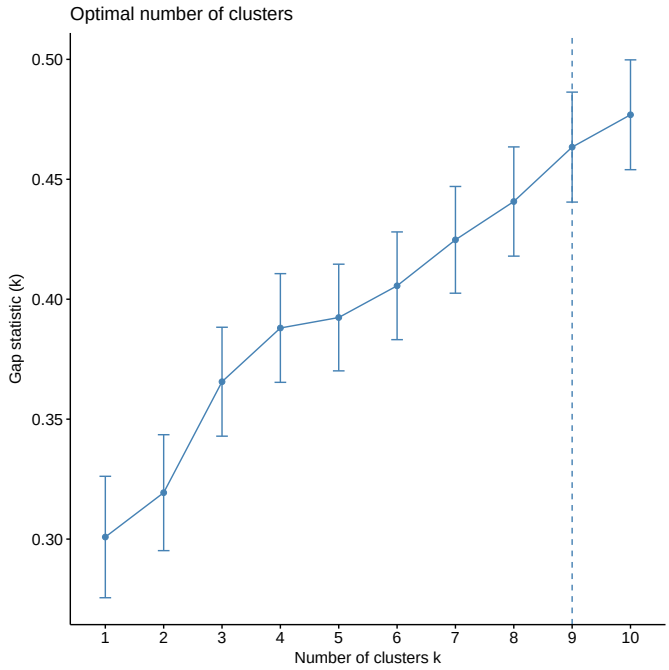


Figure 4.13: Optimal number of clusters when only the structures with experimental ΔE_{st} values measured in toluene are included. The y-axis represents the calculated gap statistic for each component. The x-axis represents the number of components.

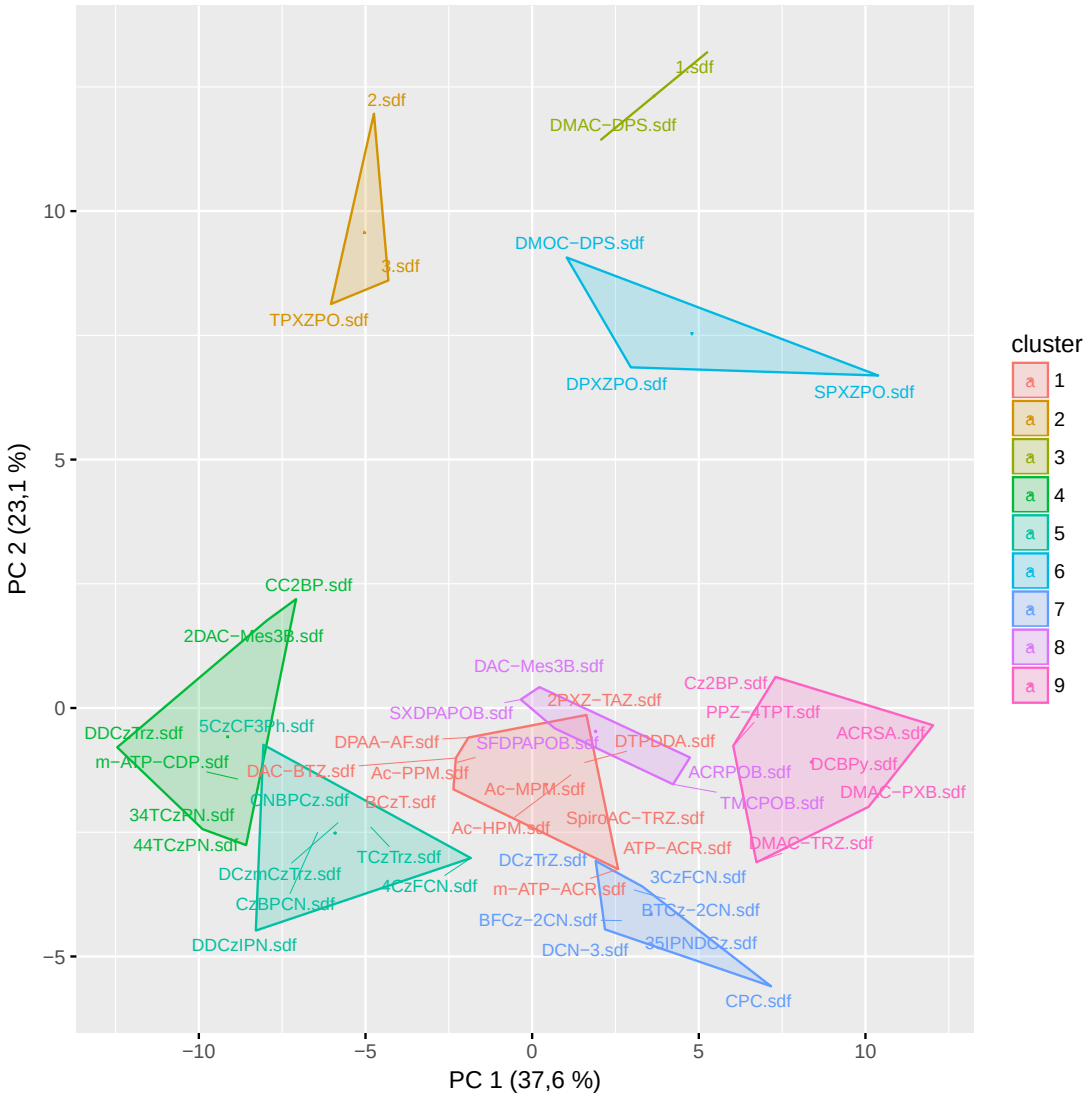


Figure 4.14: The figure display the clusters and the structures affiliation to each of these clusters. The REST set of descriptors are used and only the structures with experimental ΔE_{st} values measured in toluene are included. The y- and x-axis represents PC 2 and PC 1, respectively, with the variance explained by each PC given in parenthesis.

4.3 Partial least squares regression

Statistics from the PLSR calculation of the full data set is presented in Table 4.2, Table 4.3 and Table 4.4 for the case of no variable selection, variable selection with VIP value of 0.8 and variable selection with VIP value of 1, respectively.

Table 4.2: Results for the training and testing of the PLSR model with no variable selection. NC is the number of principal components, NV is the number of variables, R^2_{CV} is the correlation coefficient of the cross-validated training set and R^2_{test} is the correlation coefficient of the test set. Obtained during the master project, fall 2017.

			Training			Testing		
Descriptor	NV	NC	R^2_{CV}	RMSEP	MAE	R^2_{test}	RMSEP	MAE
EVA	1808	3	-0.19	0.15	0.02	0.19	0.12	0.09
EEVA	1683	1	-0.25	0.15	0.07	0.08	0.12	0.10
EVA+EEVA	3490	3	0.14	0.13	0.02	0.19	0.12	0.09
REST	58	1	0.13	0.13	0.08	0.07	0.14	0.10
EVA + REST	1871	3	-0.48	0.16	0.02	0.17	0.14	0.10
EEVA + REST	1738	1	-0.16	0.15	0.07	0.10	0.12	0.10
ALL	3554	3	0.08	0.13	0.02	0.17	0.12	0.09

Table 4.3: Results for the training and testing of the PLSR model including variable selection with VIP = 0.8. NC is the number of principal components, NV is the number of variables, R^2_{CV} is the correlation coefficient of the cross-validated training set and R^2_{test} is the correlation coefficient of the test set. Obtained during the master project, fall 2017.

			Training			Testing		
Descriptor	NV	NC	R^2_{CV}	RMSEP	MAE	R^2_{test}	RMSEP	MAE
EVA	918	3	0.54	0.09	0.02	0.19	0.12	0.09
EEVA	725	3	0.33	0.11	0.02	0.16	0.12	0.09
EVA+EEVA	1788	3	0.57	0.09	0.02	0.18	0.12	0.09
REST	35	1	0.22	0.11	0.08	0.07	0.14	0.11
EVA + REST	943	3	0.60	0.09	0.02	0.19	0.13	0.10
EEVA + REST	731	1	0.29	0.11	0.06	0.10	0.12	0.10
ALL	1823	3	0.62	0.08	0.02	0.17	0.12	0.09

Table 4.4: Results for the training and testing of the PLSR model with VIP = 1. NC is the number of principal components, NV is the number of variables, R^2_{CV} is the correlation coefficient of the cross-validated training set and R^2_{test} is the correlation coefficient of the test set.Obtained during the master project, fall 2017.

			Training			Testing		
Descriptor	NV	NC	R^2_{CV}	RMSEP	MAE	R^2_{test}	RMSEP	MAE
EVA	613	3	0.63	0.08	0.02	0.19	0.13	0.10
EEVA	537	3	0.37	0.11	0.02	0.15	0.13	0.09
EVA+EEVA	1157	3	0.59	0.09	0.02	0.16	0.12	0.09
REST	28	1	0.21	0.12	0.08	0.05	0.14	0.11
EVA + REST	623	3	0.68	0.08	0.02	0.19	0.13	0.10
EEVA + REST	537	1	0.35	0.11	0.06	0.08	0.12	0.10
ALL	1174	3	0.70	0.07	0.02	0.17	0.13	0.10

The variable selection does improve the cross-validated model, as can be seen from the increase in R^2_{CV} from 0.08 to 0.7 and the decrease in RMSEP from 0.13 to 0.07. However, it does not improve the predictive ability of the calibration model. The R^2_{test} is the same for the case of no variable selection and variable selection with VIP values of 0.8 and 1 when all the descriptors are included. The same applies to the RMSEP of the test set. Consequently, the variable selection is not justified as it should result in an increase in R^2 and a reduced RMSEP. Note that 25 % of the calibration data was ascribed as an independent test set, which corresponds to 15 objects. However, 17 objects were set aside as a test set in order to get normal distributed y-values. The remaining 43 objects were used as the training set. With a 10-fold cross-validation this amounted to each segment containing only four or five objects, which means that for each fold the cross-validated models were only tested with four or five test objects. This may have led to an overestimation of the predictive ability for the cross-validated model. None of the calculated PLSR models give any satisfactory results due to the low R^2_{test}.

For the case where all the descriptors were included, the optimal number of components, determined by the cross-validation, was three. Fig. 4.15 shows how the RMSEP improves as the number of components increase. This illustrates quite well the balance of complexity versus number of components. From the plot it can be seen that an additional component would reduce the RMSEP both for NOVARSEL and for VIP, but the reduction is considered to small compared to the increase in complexity of adding another component.

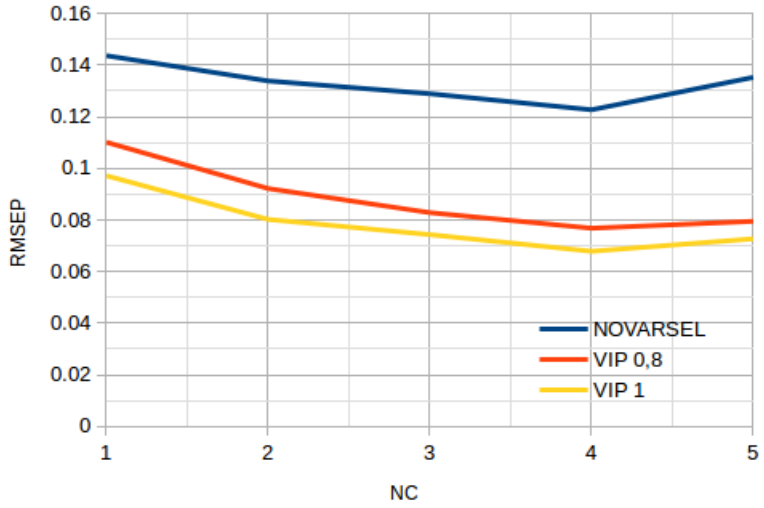


Figure 4.15: RMSEP calculated for each added component. The plot shown how the RMSEP evolve when more components are added for the case of no variable selection (NOVARSEL) and for variable selection with VIP equal 0.8 and 1.

The model fit for the training set and the response of the test set are presented in Fig. 4.16, Fig. 4.17 and Fig. 4.18 for the case where all the descriptors are included and no variable selection is performed. As can be seen, the fit of the training set is relatively good, but when an independent test set is applied to the model it reveals a quite bad predictive ability. The objects in the test set were 2, 35IPNDCz, 3CzFCN, 3, Ac-MPM, BTCz-2CN, CC2BP, Cz2BP, DAC-BTZ, DAC-Mes3B, DCzIPN, DMAC-DPS, DTC-mBPSB, DTC-pBPSB, DTPDDA, SPXZPO and TB-1PXZ. The observed values and the responses for these test objects are presented in Table 4.5. In the results from PCA, structure 2, 3, DAC-Mes3B, DMAC-DPS, DTC-mBPSB, DTC-pBPSB, SPXZPO and TB-1PXZ were observed as clusters and points separated from the rest of the population. The result shows that the structures DMAC-DPS, DTC-pBPSB, SPXZPO and TB-1PXZ have particularly bad responses.

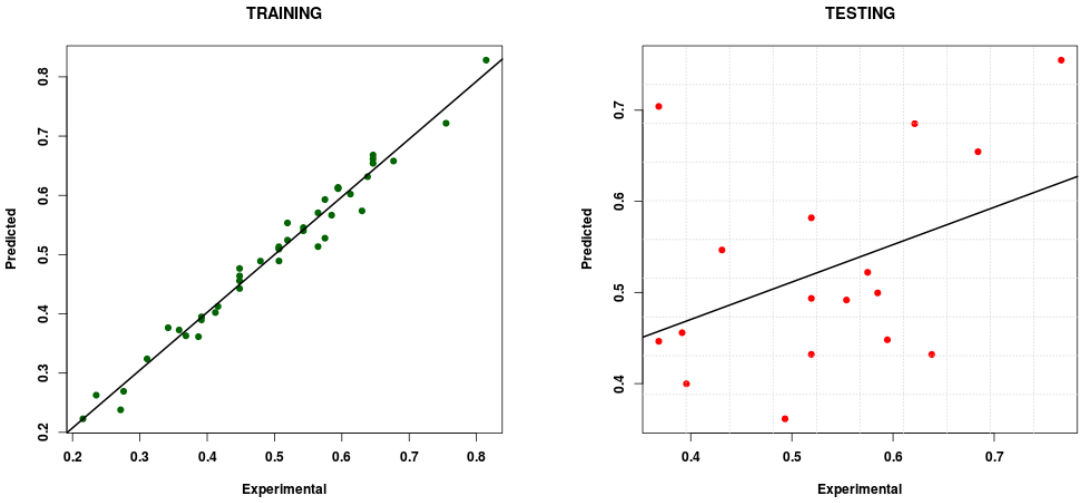


Figure 4.16: Fit of the predicted ΔE_{st} values in the training set and the testing set when all the descriptors are included. The y- and x-axis represents the predicted and the experimental ΔE_{st} values, respectively. Obtained during the master project, fall 2017.

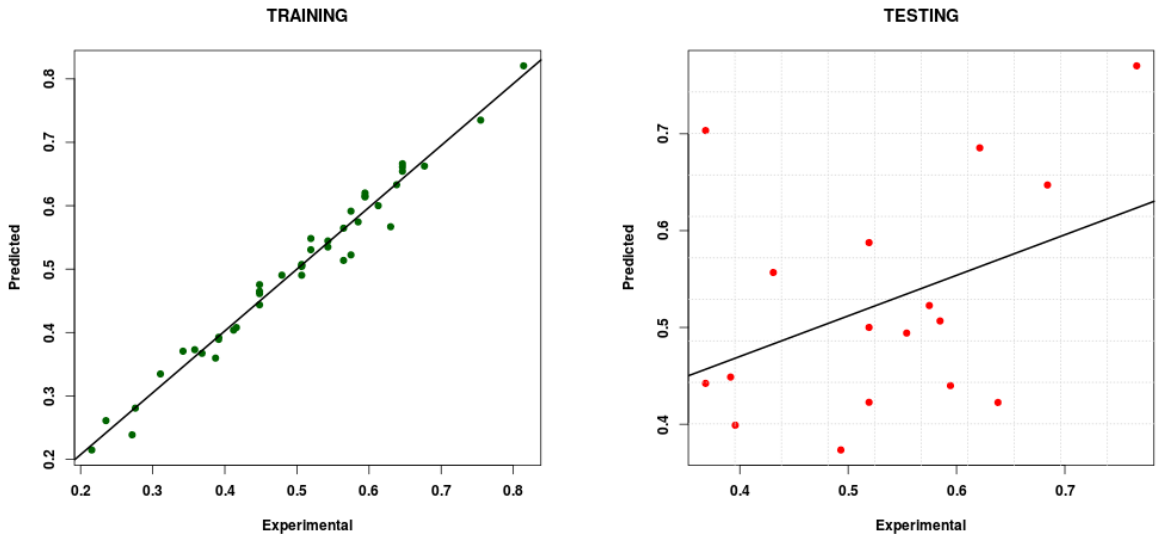


Figure 4.17: Fit of the predicted ΔE_{st} values in the training set and the testing set when all the descriptors are included and variable selection is performed with VIP=0.8. The y- and x-axis represents the predicted and the experimental ΔE_{st} values, respectively. Obtained during the master project, fall 2017.

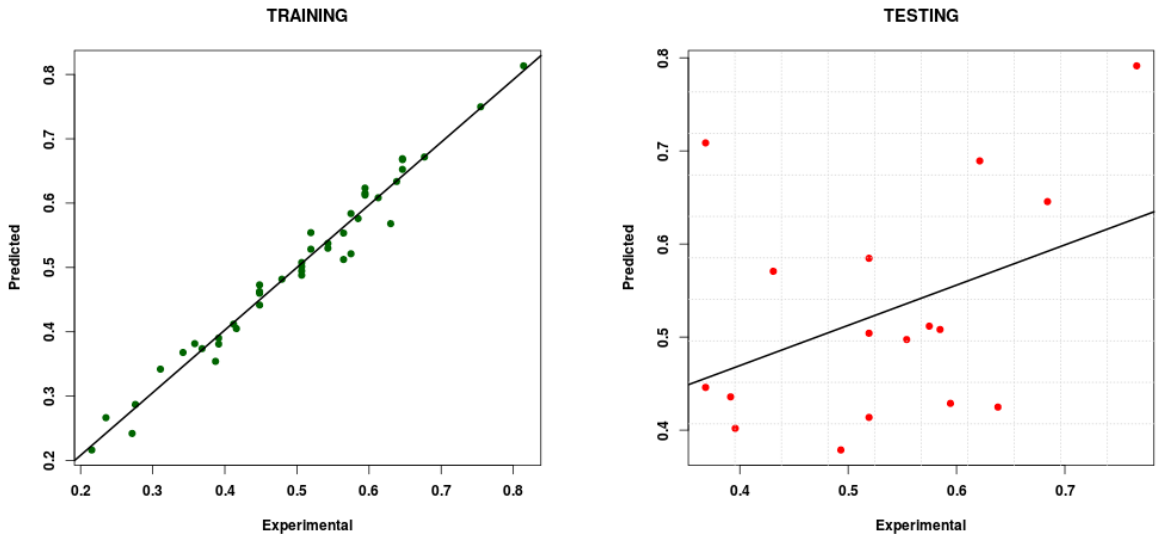


Figure 4.18: Fit of the predicted values in the training set and testing set when all the descriptors are included and variable selection is performed with VIP=1. The y- and x-axis represents the predicted and the experimental ΔE_{st} values, respectively. Obtained during the master project, fall 2017.

Table 4.5: Observed and predicted ΔE_{st} values for the test set, ΔE_{st}^{obs} and ΔE_{st}^{pred}, respectively. All descriptors are included with no variable selection. Note that the observed values are the third root of the experimental ΔE_{st} values. Obtained during the master project, fall 2017.

Structure	ΔE_{st}^{obs}	ΔE_{st}^{pred}
2	0.77	0.76
35IPNDCz	0.52	0.43
3CzFCN	0.39	0.46
3	0.68	0.65
Ac-MPM	0.58	0.52
BTCz-2CN	0.55	0.49
CC2BP	0.52	0.58
Cz2BP	0.59	0.45
DAC-BTZ	0.59	0.50
DAC-Mes3B	0.40	0.40
DCzIPN	0.37	0.45
DMAC-DPS	0.43	0.55
DTC-mBPSB	0.62	0.69
DTC-pBPSB	0.37	0.70
DTPDDA	0.52	0.49
SPXZPO	0.64	0.43
TB-1PXZ	0.49	0.36

The scores plots where all the descriptors are used without variable selection are depicted in Fig. C.1, Fig. C.2 and Fig. C.3 in Appendix C.1. It is interesting to see here that the structures with the same type of chemical scaffold are situated close to each other. For example, Ac-HPM, Ac-MPM and Ac-PPM can be observed as grouped together, which is also the case for ACRPOB, SFDPAPOB, TMCPOB and SXDPAPOB. As for the result in PCA there are some large empty areas, which means that the chemical space is not fully captured by the data. An

additional observation is that there are some objects lying relatively isolated at the periphery of the plots. These objects may be potential outliers. A plot of the studentized residuals against the leverage is presented in Fig. 4.19. As can be seen, 5CzCF3Ph, DPAA-AF, DDCzIPN, DDCzTrz and 1 have leverages beyond the leverage threshold. These points represent potential outliers in the data and should in reality be removed from the data set. However, given the poor PLSR result and the distribution of the variables and the objects in PCA and PLSR, it is strongly suspected that removing these would only result in the discovery of new ones. These five points belong to four different clusters according to Fig. 4.10, which may substantiate this claim.

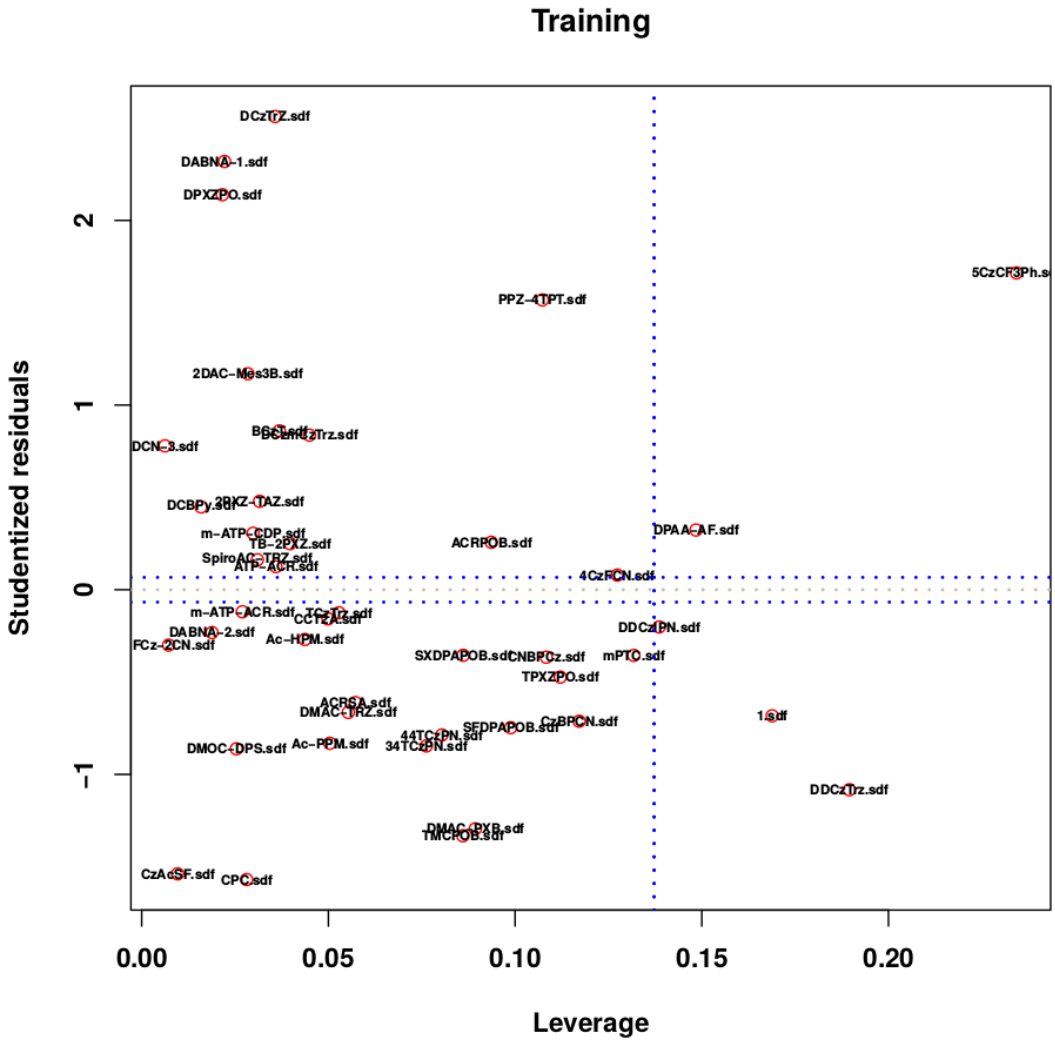


Figure 4.19: Studentized residuals plotted against leverage in the case of no variable selection. All the objects are included and all the descriptors are used.

The first impression of the structures is that all of them seem fairly rigid and that the probability that they have multiple conformations is small. Still, multiple conformations are possible and have been searched for to see if it would impact the result of the PLSR. It was found that there were five structures that had multiple conformations, namely DMAC-DPS, DMAC-TRZ, DTPDDA, SPXZPO and SpiroAC-TRZ. Two different approaches were taken to assess the conformations,

using the mean of the descriptors for each conformation pertaining to one structure and using the Boltzmann weights. The result from PLSR with and without multiple configurations is shown in Table D.1 in Appendix D. The result is shown for when all descriptors are used. By considering multiple conformations there is no significant improvement. The predictive ability of the model is still poor. The predicted values are given in Table D.2 and show no discernible change. Further study regarding conformations was therefore discarded.

4.3.1 Identification of potential outliers

An attempt to improve the PLSR model was made by removing the eleven objects which were seen in PCA as separated from the rest of the population. The statistics are given in Table 4.6 where all the descriptors are included. The model does in fact improve a little. The R^2_{CV} has increased from 0.08 for the original data to 0.16 and the R^2_{test} has increased from 0.17 to 0.46. Although this is promising, the RMSEP of the training set is approximately the same and the reduction of RMSEP for the test set from 0.12 to 0.09 is negligible. A key point to mention here is that by removing the eleven objects the test set has naturally changed from that of the original PLSR model. As a consequence, these two models are not directly comparable. The test set consisted of the structures 35IPNDCz, 3CzFCN, Ac-MPM, CNBPCz, CPC, Cz2BP, CzBPCN, DABNA-1, DCzIPN, m-ATP-ACR, SFDPAPOB and SXDPAPOB. The predicted values of the test set are given in Table C.1 in Appendix C.2 when all descriptors are included and without variable selection. The model fit of the training set and the test set is given in Fig. C.4, Fig. C.5 and Fig. C.6 in Appendix C.2 without variable selection and variable selection with VIP values 0.8 and 1, respectively. Another important factor to consider here is that the data set now only consists of 49 structures of which 12 are reserved as a test set. The number of objects in the test set should probably have been reduced a little, but it would not have changed the fact that 49 objects are a very small data set to perform regression on. The validity of the result is therefore compromised.

Table 4.6: Results for the training and testing of the PLSR model for which the eleven potential outliers are excluded. All descriptors are included. NC is the number of principal components, NV is the number of variables, R^2_{CV} is the correlation coefficient of the cross-validated training set and R^2_{test} is the correlation coefficient of the test set.

			Training			Testing		
Descriptor	NV	NC	R^2_{CV}	RMSEP	MAE	R^2_{test}	RMSEP	MAE
NOVARSEL	3523	3	0.16	0.12	0.01	0.46	0.09	0.08
VIP = 0.8	1822	4	0.70	0.07	0.007	0.49	0.09	0.08
VIP = 1	1194	3	0.73	0.07	0.01	0.41	0.09	0.08

The scores plots, given in Fig. C.7, Fig. C.8 and Fig. C.9 in Appendix C.2, show the same trend as that of the original data. Fairly isolated objects can be seen at the periphery of the plots. The studentized residuals is plotted against the leverage in Fig. 4.20. There are still potential outliers present, where four objects can be seen having a larger leverage than the threshold. The reason for not removing these outliers are the same as before, it is suspected that removing them would yield new outliers. Moreover, the small data set containing objects which are highly heterogeneous, increases the possibility of new outliers being discovered when others are removed. Judging by the improvement in R^2_{test} the model seems at first glace significantly better than that of the original data. But considering the small data set, small reduction in RMSEP and the fact that there is still potential outliers present the result is still unsatisfactory.

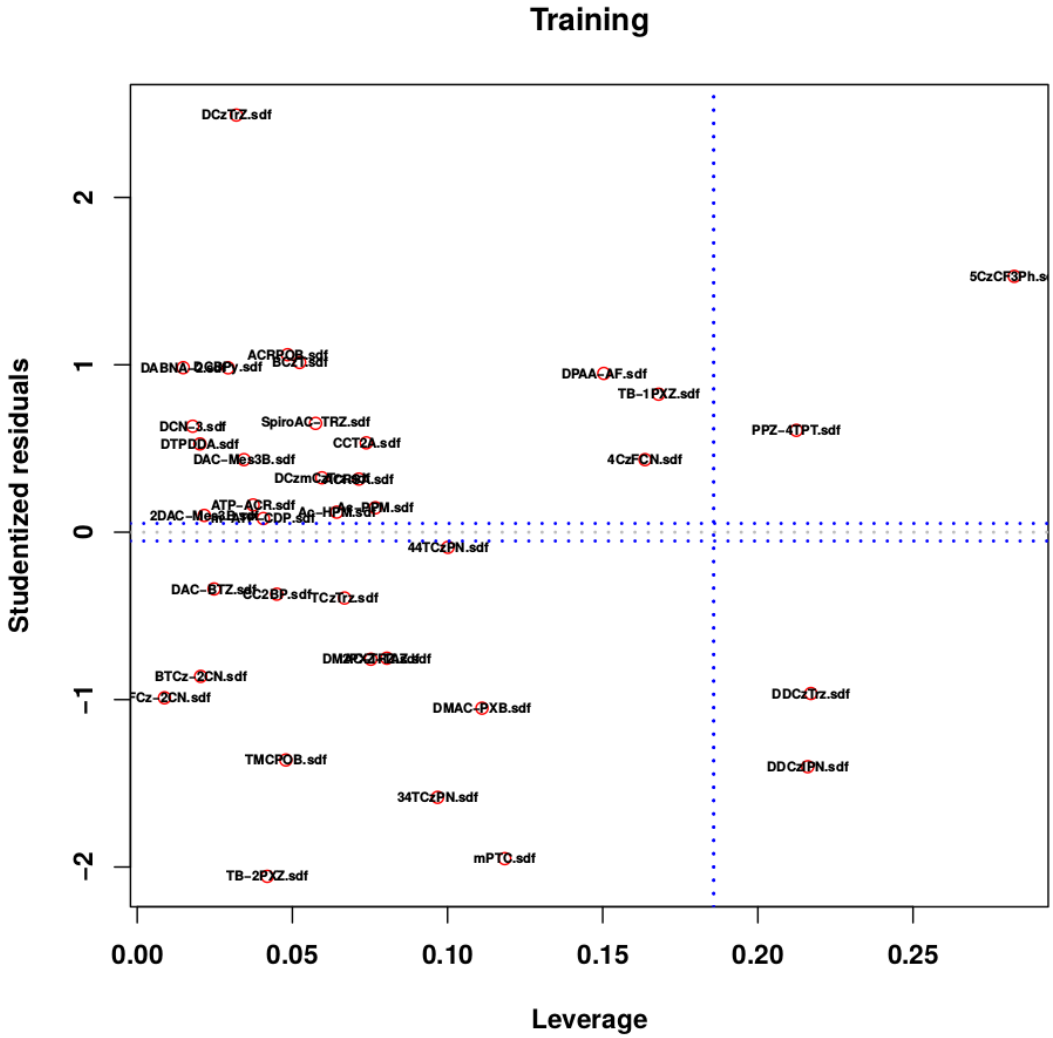


Figure 4.20: Studentized residuals plotted against leverage in the case where the eleven potential outliers are removed. No variable selection is performed and all descriptors are included.

4.3.2 Solvent effects

Another approach to improve the PLSR result was to only consider structures for which the reported experimental ΔE_{st} value was measured in toluene. In that way, solvent effects would to some extent be accounted for. The statistics are presented in Table 4.7. The R^2_{test} improves slightly, but the RMSEP of the test set is approximately the same as the result when the full data set is used. Notice that without variable selection the optimal number of components is 10. The RMSEP plot without variable selection is given in Fig. 4.21. The decrease in RMSEP for each added component is very small. Only when going from 6 to 7 or from 9 to 10 components is the decrease more than 5 %. Considering the model fit of the training set and test set given in Fig. C.10 in Appendix C.3 in conjunction with the change in RMSEP, the result clearly indicates that the model is overfitted. The fit of the training set is exceptionally good, at least relative to that seen for the other models. The fit of the test set however shows a poor predictive ability. The overfitting may be a symptom of the small data set. By removing the structures which did not have experimental ΔE_{st} values measured in toluene, only 50 structures remained. The studentized residual plotted against the leverage is given in Fig. 4.22 and it reveals five potential outliers. These points belong to four different cluster according to Fig. 4.14.

Table 4.7: Results for the training and testing of the PLSR model where only the structures with experimental ΔE_{st} values measured in toluene are considered. All descriptors are included. NC is the number of principal components, NV is the number of variables, R^2_{CV} is the correlation coefficient of the cross-validated training set and R^2_{test} is the correlation coefficient of the test set.

			Training			Testing		
Descriptor	NV	NC	R^2_{CV}	RMSEP	MAE	R^2_{test}	RMSEP	MAE
NOVARSEL	3520	10	0.21	0.11	0.0001	0.28	0.13	0.09
VIP = 0.8	1905	3	0.54	0.08	0.02	0.22	0.14	0.10
VIP = 1	1242	3	0.66	0.07	0.02	0.26	0.14	0.10

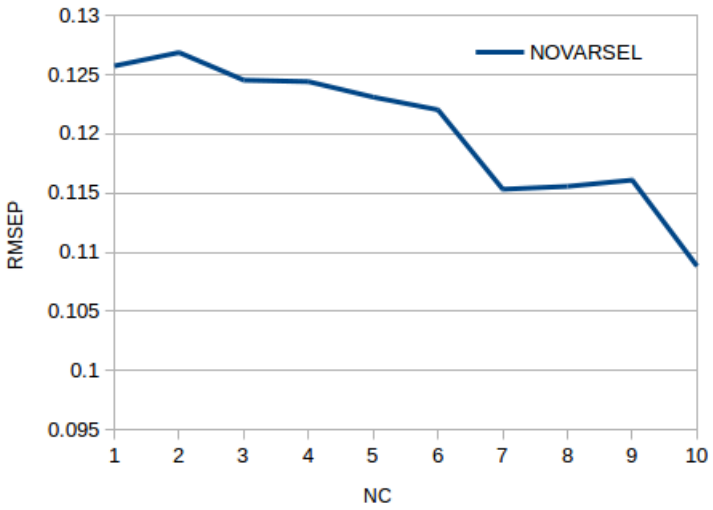


Figure 4.21: RMSEP calculated for each added component. The plot shows how the RMSEP evolve when more components are added. No variable selection (NOVARSEL) is performed and only the structures with experimental ΔE_{st} values measured in toluene are considered..

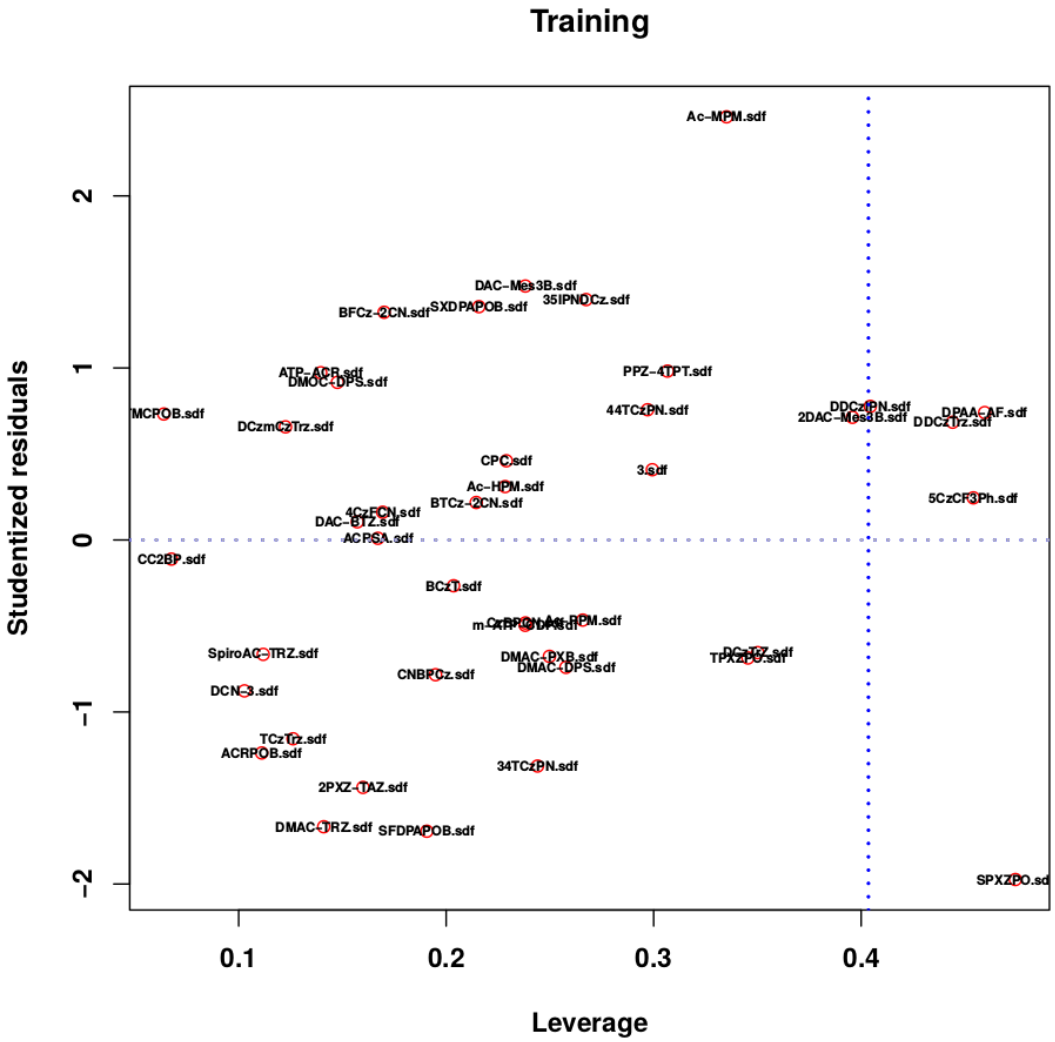


Figure 4.22: Studentized residuals plotted against leverage for the case where only objects with experimental ΔE_{st} values measured in toluene are considered. No variable selection is performed and all the descriptors are included.

4.4 Cubist

The result from the Cubist model is given in Table 4.8 when the REST set of descriptors are included. And the most attributing variables to the model is given in Table 4.9. The model was constructed using 50 committees and 9 neighbours. Both the correlation coefficient R^2_{CV} and the error have improved from the PLSR, which suggest that there is most likely a non-linear relationship between the chosen descriptors and ΔE_{st}. The predicted ΔE_{st} values for each structure are given in Table E.1 in Appendix E. The Spearman's rank correlation coefficient between the predicted and observed values is 0.98, which means that there is an almost perfect positive monotonic relationship between the observed and the predicted values. The correlation coefficieent between the predicted and observed values was also calculated, which gave a R^2 value of 0.96. Notice however that here the same objects are used to construct the model and for testing. This is not the same as for PLSR where an independent test set is used. Thus, R^2 and R^2_{test} is not the same. The high R^2 value may be misleading and is not sufficient to determine the models predictive ability. Moreover, the interest here is to see how the model respond to new unseen samples, which is the core objective of this thesis.

Table 4.8: Performance metrics for the Cubist model when all the objects are included and the REST set of descriptors is used. NOVARSEL denotes the results obtained when no variable selection is performed and VARSEL denotes results obtained when variable selection is performed. ND denotes the number of descriptors included in the model, R^2_{CV} is the correlation coefficient of the cross-validated model and RMSE is the root mean squared error.

	ND	R^2_{CV}	RMSE	MAE
NOVARSEL	86	0.38	0.10	0.079
VARSEL	43	0.48	0.08	0.068

An interesting aspect of the most contributing descriptors in the Cubist model is the introduction of shape descriptors. In the PCA calculations only CPSA and MOPAC descriptors was seen as contributing to the first three PCs. In PCA the objective is to explain as much as possible of the variance in the descriptor data. It is not related to ΔE_{st} at all. The fact that different descriptors are shown most contributing to the model when ΔE_{st} is actively used in the model construction may give some indication as to which descriptors are relevant for ΔE_{st}. As CPSA is contributing to a large extent in both PCA and Cubist, it is reasonably to assume that these descriptors are important for ΔE_{st} and are strong distinguishing factors between the structures. Generalizing the CPSA descriptors it can be said that the charge distribution is an important molecular representation of the structures in order to relate them to ΔE_{st}. The occurrence of the autocorrelated charge properties, total dipole moment and charge dipole moment as most contributing descriptors also indicates that charge distribution is in fact an important representation. It is hard to say if the shape descriptors are directly relevant for predicting ΔE_{st} based on just the attribution in Cubist. It may perhaps be a consequence of the data containing structures which vary a lot in size and shape.

Table 4.9: Descriptor attribution in the Cubist model when all the structures are included and the REST set of descriptors is used.

Percentage	Descriptor
50%	Autocorrelated charge property 8
30%	CPSA-RNCS
40%	Ovality
20%	Radius of gyration
20%	CPSA-FNSA-2
17%	CPSA-WNSA-3
10%	CPSA-PNSA-5
10%	Total dipole moment
10%	Maximum DER
10%	Minimum DNR
10%	CPSA-PNSA-1
10%	Molecular eccentricity
10%	Electrophilicity
10%	Asphericity
10%	Minimum DER
10%	Maximum SPOL
10%	CPSA-PNSA-2
10%	CPSA-RPCG
10%	Heat of formation
10%	CPSA-WNSA-2
10%	Maximum DNR
7%	Charge dipole moment
7%	CPSA-PPSA-5
7%	CPSA-FNSA-3
6%	Autocorrelated charge property 7
4%	CPSA-PPSA-4

4.4.1 Identification of potential outliers

Here the results of the Cubist modelling are presented when the eleven points seen in PCA and k-means cluster analysis as separated from the rest of the population is removed. The reason for removing these objects is the same as that for the PLSR case, to see if that would improve the Cubist model. There is a large area between the separated data and the rest of the objects where the chemistry is not spanned. The goal is to reduce this non-spanned area by removing these eleven objects. The results from the Cubist model is presented in Table 4.10 and the most attributing descriptors are listed in Table 4.11. The model was constructed using 10 committees and 7 neighbours. Without variable selection, the correlation coefficient and the error have slightly improved. With variable selection the result is approximately the same as for the previous model. Although an improvement can be observed the model is still not accurate enough to consider it a good model.

Table 4.10: Performance metrics for the Cubist model when the eleven potential outliers have been removed and the REST set of descriptors is used. NOVARSEL denotes the results obtained when no variable selection is performed and VARSEL denotes results obtained when variable selection is performed. ND denotes the number of descriptors included in the model, R^2_{CV} is the correlation coefficient of the cross-validated model and RMSE is the root mean squared error.

	ND	R^2_{CV}	RMSE	MAE
NOVARSEL	86	0.44	0.08	0.061
VARSEL	36	0.48	0.07	0.057

Much of the same types of descriptors can be seen as most contributing. According to the eleven points having on average a higher dipole moment, it not surprising that the total dipole moment descriptor is no longer a contributing descriptor. However, the shape of the structures still seem to be an important distinguishing factor. The same can be said for the different autocorrelated charge descriptors and the CPSA descriptors.

Table 4.11: Descriptor attribution in the Cubist model when the eleven potential outliers are excluded and the REST set of descriptors is used

Percentage	Descriptor
20 %	Maximum DNR
32 %	Ovality
30 %	Maximum DER
23 %	Globularity
22 %	Autocorrelated charge property 4
15 %	CPSA-PNSA-4
15 %	CPSA-WNSA-3
12 %	Autocorrelated charge property 2
10 %	Spherosity
10 %	Maximum SPOL
10 %	CPSA-FNSA-1
5 %	CPSA-PNSA-3
5 %	Molecular eccentricity
5 %	CPSA-PNSA-1
5 %	CPSA-DPSA-3

The predicted values from this model is given in Table E.1 in Appendix E. R^2 is in this case calculated to be 0.98. It is a slight increase from the previous model, but again, this is not a sufficient measure of the models predictive power as the prediction is not done on an independent test set. Calculating the Spearman's rank correlation gives a value of 0.99, meaning an almost perfect positive monotonic relationship between the predicted values and the observed values.

4.4.2 Solvent effects

Here the solvent effects are considered. Note that it is only considered in the sense that only the structures with experimental values of ΔE_{st} measured in toluene are included. The solvent effects are not included in the calculation of the descriptors. The results from the Cubist model with 50 committees and 5 neighbours used

are given in Table 4.12. The most attributing descriptors are given in Table 4.13. The correlation coefficient show some improvement from the Cubist model where all the objects are included. This applies both to when variable selection is not performed and when variable is performed. The RMSE has improved slightly for when no variable selection is performed, whereas the RMSE is the same for when variable selection is performed. The predicted values for this model are given in Table E.1 in Appendix E. An R^2 of 0.98 was calculated, but the same notes made earlier about the R^2 applies here. The Spearman's rank correlation coefficient for this case is 0.995.

Table 4.12: Performance metrics for the Cubist model when only the objects with experimental ΔE_{st} values measured in toluene are included and the REST set of descriptors is used. NOVARSEL denotes the results obtained when no variable selection is performed and VARSEL denotes results obtained when variable selection is performed. ND denotes the number of descriptors included in the model, R^2_{CV} is the correlation coefficient of the cross-validated model and RMSE is the root mean squared error.

	ND	R^2_{CV}	RMSE	MAE
NOVARSEL	86	0.51	0.08	0.07
VARSEL	12	0.59	0.08	0.07

Much of the same type of descriptors can be seen as most contributing to the model. The total dipole moment has reentered as contributing descriptor. As seen in PCA and k-means cluster analysis, only considering objects with experimental ΔE_{st} values measured in toluene did not remove the separation of the data, just reduced the number of objects in the separated group. Therefor, some of the objects which on average have a larger dipole moment are still included, explaining why the dipole moment is again a contributing descriptor.

Table 4.13: Descriptor attribution in the Cubist model when only the objects with experimental ΔE_{st} values measured in toluene are included and the REST set of descriptors is used.

Percentage	Descriptor
34 %	Charge dipole moment
29 %	CPSA-PNSA-3
12 %	Total dipole moment
18 %	Asphericity
48 %	Maximum DER
34 %	Ovality
32 %	Autocorrelated charge property 8
30 %	Spherosity
20 %	CPSA-PNSA-4
15 %	Autocorrelated charge property 6
15 %	CPSA-WNSA-3
15 %	LUMO energy
14 %	CPSA-PNSA-1
14 %	Radius of gyration
14 %	Minimum DER
9 %	Autocorrelated charge property 2
8 %	Minimum DNR
5 %	Molecular eccentricity
4 %	Autocorrelated charge property 4

4.5 Density functional theory

As explained in the methodology section (Section 2.4), the quantum chemical computations where performed in three steps on a selected set of structures. The reason for selecting the structures in the way presented in Section 2.4 was not just in regards to the computational time, but also to make sure that the structures represented the whole chemical space spanned by the data. Unfortunately, we were not successful in completing the TDDFT computations within the time frame of this master thesis. There were some issue with the TDDFT calculation causing it to produce excited state energies with severe discrepancies. Some of the articles on the structures used in this thesis have reported ΔE_{st} values calculated with TDDFT. An example is the structures 1, 2 and 3 [26]. Zang et al. reported experimentally determined values for these structures of 0.54, 0.45 and 0.32 eV, respectively. The ΔE_{st} values they calculated using TDDFT with a B3LYP functional and 6-31G* basis set was however 0.65, 0.6 and 0.34 eV, respectively. This goes to show that calculating high accuracy excited stated is a difficult task.

We have sadly not been able to find the source of the TDDFT problem. The output coordinates from the DFT calculation with the COSMO solvation model are therefore given in Appendix F so that it can be used in further study. The chosen basis set and functional have to be assessed in order to determine if they are suited for the calculation of the excited state energies of the structures in our data set. The computation time for the geometry optimization in solvent took on average 118 hours, with the longest computation running for 323 hours. The TDDFT computations that did converge included only the three smallest structures, for which the computation time was on average 44 hours. In comparison, PLSR and Cubist only takes a couple of minutes to complete.

5 Conclusion

In this thesis a QSPR approach was used in an attempt to create a model which could aid the design of blue TADF based OLEDs. The structural data and experimental ΔE_{st} values for 60 different blue emitting structures used in this work was collected from a recent review paper. Based on different molecular representations of these structures, a set of descriptors were calculated. In the exploratory analysis of these descriptors the PCA and k-means cluster analysis was performed to investigate the variance in the data, patterns and potential outliers. Both methods revealed a highly uneven distribution of objects and large areas where the chemical space was not-spanned. Eleven objects were seen as separated from the rest of the population. In the k-means cluster analysis the objects were separated into nine different cluster. These observation suggested that the data was fairly heterogeneous, differing in multiple ways. All the structures consisted of a variety of different donor and acceptor moieties and with a multitude of different donor and acceptor architectures. They also differed in shape, size and charge distribution.

In the regression analysis the linear method of PLSR was performed first. When all the objects were included the best model had a R^2_{CV} of 0.7 and a R^2_{test} of 0.17. Two different approaches to improve the PLSR was tested. The first was to remove the eleven objects seen in PCA and k-means cluster analysis as separated from the rest of the population. The second approach was to only consider objects which were reported having experimental ΔE_{st} values measured in toluene. The best result from the first approach was with a R^2_{CV} of 0.73 and R^2_{test} of 0.41 when variable selection was performed with VIP=1. The best result from the second approach was also obtain when variable selection was performed with VIP=1, yielding a R^2_{CV} of 0.66 and a R^2_{test} of 0.26. However, removing objects from the data set may have caused the data set to be too small for the purpose of a PLSR. The performance of these two models were therefore questioned despite the improvement in the correlation coefficients. The relatively large difference in R^2_{CV} and R^2_{test} for these models suggested an overfitting of the data, especially in the case where the solvent effects were considered. The RMSEP, both for the training set and the test set, was approximately the same for all the PLSR models. All put together, the PLSR

did not give satisfactory results.

The non-linear regression tree method Cubist was performed for which the results showed an improvement from the PLSR models. The Cubist model with all the objects included had a R^2_{CV} of 0.38 without variable selection. With variable selection this result was improved giving a R^2_{CV} of 0.48. A small reduction in RMSE could also be observed when variable selection was performed. The same two approaches was tested here as for the PLSR to see if it would improve the model. When the eleven potential outliers were removed the result displayed a small improvement for the model without variable selection, but approximately the same result was obtained when variable selection was performed. Considering only objects which had experimental ΔE_{st} values measured in toluene improved the model even more. The result showed a R^2_{CV} of 0.51 and 0.59 without variable selection and when variable selection was performed, respectively. The RMSE was approximately the same as the best result from the two other models. Predictions were done with all the Cubist models, but regrettably not with an independent test sets. The same objects used for constructing the models were used in the testing of these models. The predictive ability of the Cubist models was therefor not tested properly. The writer acknowledges this flaw and that and independent test should have been employed to each Cubist model. That being said, the relative small R^2_{CV} indicates that the models would not be able to display a predictive ability of a satisfactory level of R^2_{test} above at least 0.8.

Non of the models created in this thesis have shown particularly good predictive abilities. There are different aspects of the data used in this study that may have caused these failings. The data have been shown to be quite heterogeneous. The small data set has therefor been rendered insufficient for spanning the whole chemical space. All the different and varying chemical properties of the blue emitting structures are therefore not fully captured. A consequence of ΔE_{st} not having been used as a property of interest in a QSRP study before is that the molecular descriptors most relevant for predicting ΔE_{st} are unknown. The most attributing descriptors in the Cubist models may give some indications as to which molecular representations may be relevant, in particular the charge distribution. However,

more research on molecular descriptors relevant for ΔE_{st} must be performed in order to say anything certain. Another factor which may have caused the poor performing models is ΔE_{st} itself. Although it is a critical property for achieving TADF, which is the reason it is chosen in this study, it may not be the easiest property of interest to use in a QSPR approach. ΔE_{st} has a fairly small range, which requires the models to be highly accurate in order to give good responses. Solvent effects may also affect the ΔE_{st}. In this study the solvent effects are only considered indirectly by generating models using only objects with experimental ΔE_{st} values measured in toluene. Ideally the molecular descriptors should be calculated using a solvation model for explicitly including the solvent effects. Other researchers have been successful with using the emission maxima and the glass transition temperature in QSRP approaches.

A TDDFT computation for determining ΔE_{st} for a selected set of structures was attempted using the CAM-B3LYP functional and the aug-pcs-1 basis set. The calculations that did converge showed serious discrepancies. The difficulty of obtaining accurate results for the excited state energies using TDDFT was exemplified with an example from Zhang and coworkers' previous work done one three of the structures used in this study. The functional and basis set used have to be assessed in order to determine their applicability in these calculations.

There are several different ways to improve the result presented in this study. The first thing is that the data set should be increased as much as possible in order to properly span the chemical space. If possible, a data set comprised of structures with the same type of scaffold is recommended. The choice of molecular representation of the structures can also be changed in order to achieve molecular descriptors which may be more suited for the prediction of ΔE_{st}.

References

[1] C. W. Tang, S. A. Van Slyke. Organic electroluminescent diodes. *Applied Physics Letters*, 51(12):913–915, 1987.

[2] M. Y. Wong, E. Z. -Colman. Purely organic thermally activated delayed fluorescence materials for organic light-emitting diodes. *Advanced Materials*, 29(22):1605444–n/a, 2017.

[3] Braemac. Oled applications. `http://www.braemac.co.uk/oledapplications.html`. Accessed: 11.01.2018.

[4] Z. Yang, Z. Mao, Z. Xie, Y. Zhang, S. Liu, J. Zhao, J. Xu, Z. Chi, M. P. Aldred. Recent advances in organic thermally activated delayed fluorescence materials. *Chemical Society Review*, 46:915–1016, 2017.

[5] A. Buckley. *Organic Light-Emitting Diodes (OLEDs): Materials, Devices and Applications*. Elsevier Science, 2013.

[6] C. Freudenrich (How Stuff Works). How oleds work. `https://electronics.howstuffworks.com/oled1.htm`, 2018. Accessed: 11.01.2018.

[7] M. Oehzelt, K. Akaike, N. Koch, G. Heimel. Energy-level alignment at organic heterointerfaces. *Science Advances*, 1(10), 2015.

[8] J. Kalinowski. *Organic Light-Emitting Diodes: Principles, Characteristics and Processes*. Marcel Dekker, New York, 2005.

[9] K. S. Yook, J. Y. Lee. Organic materials for deep blue phosphorescent organic light-emitting diodes. *Advanced Materials*, 24(24):3169–3190, 2012.

[10] W. Li, D. Liu, F. Shen, D. Ma, Z. Wang, T. Feng, Y. Xu, B. Yang, Y. Ma. A twisting donor-acceptor molecule with an intercrossed excited state for highly efficient, deep-blue electroluminescence. *Advanced Functional Materials*, 22(13):2797–2803, 2012.

[11] P. Zhang, W. Dou, Z. Ju, L. Yang, X. Tang, W. Liu, Y. Wu. A 9,9′-bianthracene-cored molecule enjoying twisted intramolecular charge transfer to enhance radiative-excitons generation for highly efficient deep-blue oleds. *Organic Electronics*, 14(3):915 – 925, 2013.

[12] Y. Lu. *Solitons & Polarons in Conducting Polymers*. World Scientific, Singapore, 1965.

[13] J. H. Burroughes, D. D. C. Bradley, A. R. Brown, R. N. Marks, K. Mackay, R. H. Friend, P. L. Burns, A. B. Holmes. Light-emitting diodes based on conjugated polymers. *Nature*, 347:539–541, 1990.

[14] R. H. Friend, R. W. Gymer, A. B. Holmes, J. H. Burroughes, R. N. Marks, C. Taliani, D. D. C. Bradley, D. A. Dos Santos, J. L. Brédas, M. Lögdlund, W. R. Salaneck. Electroluminescence in conjugated polymers. *Nature*, 397:121–128, 1999.

[15] A. Endo, K. Sato, K. Yoshimura, T. Kai, A. Kawada, H. Miyazaki, C. Adachi. Efficient up-conversion of triplet excitons into a singlet state and its application for organic light emitting diodes. *Applied Physics Letters*, 98(083302), 2011.

[16] Y. Tao, K. Yuan, T. Chen, P. Xu, H. Li, R. Chen, C. Zheng, L. Zhang, W. Huang. Thermally activated delayed fluorescence materials towards the breakthrough of organoelectronics. *Advances Materials*, 26:7931–7958, 2014.

[17] P. Atkins, R. Friedman. *Molecular Quantum Mechanics*. Oxford University Press Inc., New York, 2011.

[18] C. Nantasenamat, C. Isarankura-Na-Ayudhya, N. ansila, T. Naenna, V. Prachayasittikul. Prediction of gfp spectral properties using artificial neural network. *Journal of Computational Chemistry*, 28(7):1275–1289, 2007.

[19] C. -H. Chen, K. Tanaka, K. Funatsu. Random forest approach to qspr study of fluorescence properties combining quantum chemical descriptors and solvent conditions. *Journal of Fluorescence*, 28(2):695–706, Mar 2018.

[20] R. Barbosa-da-Silva, R. Stefani . Qspr based on support vector machines to predict the glass transition temperature of compounds used in manufacturing oleds. *Molecular Simulation*, 39(3):234–244, 2013.

[21] J. Gasteiger, editor. *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*. Wiley-VCH, Weinheim, 2003.

[22] Oprea Tudor I. *Chemoinformatics and the Quest for Leads in Drug Discovery*, chapter 4.1, pages 1508–1531. Wiley-Blackwell, 2008.

[23] R. Todeschini, V. Consonni. *Molecular Descriptors for Chemoinformatics Volume I: Alphabetical Listing / Volume II: Appendices, References*. Methods & Principles in Medicinal Chemistry. Wiley-VCH, 2nd ed., rev. and enl. edition, 2009.

[24] ChemAxon. Marvinsketch (version 17.22.0).

[25] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011.

[26] Q. Zhang, J. Li, K. Shizu, S. Huang, S. Hirata, H. Miyazaki, C. Adachi. Design of efficient thermally activated delayed fluorescence materials for pure blue organic light emitting diodes. *Journal of the American Chemical Society*, 134(36):14706–14709, 2012.

[27] K. Suzuki, S. Kubo, K. Shizu, T. Fukushima, A. Wakamiya, Y. Murata, C. Adachi, H. Kaji. Triarylboron-based fluorescent organic light-emitting diodes with external quantum efficiencies exceeding 20 %. *Angewandte Chemie International Edition*, 54(50):15231–15235, 2015.

[28] J. Lee, K. Shizu, H. Tanaka, H. Nomura, T. Yasuda, C. Adachi. Oxadiazole- and triazole-based highly-efficient thermally activated delayed fluorescence emitters for organic light-emitting diodes. *Journal of Materials Chemistry C*, 1(30):4599–4604, 2013.

[29] M. Kim, S. K. Jeon, S. H. Hwang, S. S. Lee, E. Yu, J. Y. Lee. Highly efficient and color tunable thermally activated delayed fluorescent emitters using a "twin emitter" molecular design. *Chemical Communications*, 52(2):339–342, 2016.

[30] B. Li, H. Nomura, H. Miyazaki, Q. Zhang, K. Yoshida, Y. Suzuma, A. Orita, J. Otera, C. Adachi. Dicarbazolyldicyanobenzenes as thermally activated delayed fluorescence emitters: Effect of substitution position on photo luminescent and electroluminescent properties. *Chemistry Letters*, 43(3):319–321, 2014.

[31] Y. J. Cho, B. D. Chin, S. K. Jeon, J. Y. Lee. 20% external quantum efficiency in solution-processed blue thermally activated delayed fluorescent devices. *Advanced Functional Materials*, 25(43):6786–6792, 2015.

[32] L. Mei, J. Hu, X. Cao, F. Wang, C. Zheng, Y. Tao, X. Zhang, W. Huang. The inductive-effect of electron withdrawing trifluoromethyl for thermally activated delayed fluorescence: tunable emission from tetra- to penta-carbazole in solution processed blue oleds. *Chemical Communications*, 51(65):13024–13027, 2015.

[33] R. Komatsu, H. Sasabe, Y. Seino, K. Nakao, J. Kido. Light-blue thermally activated delayed fluorescent emitters realizing a high external quantum efficiency of 25% and unprecedented low drive voltages in oleds. *Journal of Materials Chemistry C*, 4(12):2274–2278, 2016.

[34] M. Numata, T. Yasuda, C. Adachi. High efficiency pure blue thermally activated delayed fluorescence molecules having 10h-phenoxaborin and acridan units. *Chemical Communications*, 51(46):9443–9446, 2015.

[35] K. Nasu, T. Nakagawa, H. Nomura, C. J. Lin, C. H. Cheng, M. R. Tseng, T. Yasuda, C. Adachi. A highly luminescent spiro-anthracenone-based organic light-emitting diode exhibiting thermally activated delayed fluorescence. *Chemical Communications*, 49(88):10385–10387, 2013.

[36] T. Takahashi, K. Shizu, T. Yasuda, K. Togashi, A. Chihaya. Donor-acceptor-structured 1,4-diazatriphenylene derivatives exhibiting thermally activated delayed fluorescence: design and synthesis, photophysical properties and oled characteristics. *Science and Technology of Advanced Materials*, 15(3):034202, 2014.

[37] K. Shizu, H. Noda, H. Tanaka, M. Taneda, M. Uejima, T. Sato, K. Tanaka, H. Kaji, C. Adachi. Highly efficient blue electroluminescence using delayed-fluorescence emitters with large overlap density between luminescent and ground states. *The Journal of Physical Chemistry C*, 119(47):26283–26289, 2015.

[38] D. R. Lee, S. H. Hwang, S. K. Jeon, C. W. Lee, J. Y. Lee. Benzofurocarbazole and benzothienocarbazole as donors for improved quantum efficiency in blue thermally activated delayed fluorescent devices. *Chemical Communications*, 51(38):8105–8107, 2015.

[39] S. Y. Lee, T. Yasuda, Y. S. Yang, Q. Zhang, C. Adachi. Luminous butterflies: Efficient exciton harvesting by benzophenone derivatives for full-color delayed fluorescence oleds. *Angewandte Chemie International Edition*, 53(25):6402–6406, 2014.

[40] C. Mayr, S. Y. Lee, T. D. Schmidt, T. Yasuda, C. Adachi, W. Brütting. Efficiency enhancement of organic light-emitting diodes incorporating a highly oriented thermally activated delayed fluorescence emitter. *Advanced Functional Materials*, 24(33):5232–5239, 2014.

[41] Y. J. Cho, S. K. Jeon, S.-S. Lee, E. Yu, J. Y. Lee. Donor interlocked molecular design for fluorescence-like narrow emission in deep blue thermally activated delayed fluorescent emitters. *Chemistry of Materials*, 28(15):5400–5405, 2016.

[42] W. Liu, C. J. Zheng, K. Wang, Z. Chen, D. Y. Chen, F. Li, X. M. Ou, Y. P. Dong, X. H. Zhang. Novel carbazol-pyridine-carbonitrile derivative

as excellent blue thermally activated delayed fluorescence emitter for highly efficient organic light-emitting devices. *ACS Applied Materials & Interfaces*, 7(34):18930–18936, 2015.

[43] I. H. Lee, W. Song, J. Y. Lee, S.-H. Hwang. High efficiency blue fluorescent organic light-emitting diodes using a conventional blue fluorescent emitter. *Journal of Materials Chemistry C*, 3(34):8834–8838, 2015.

[44] T. Hatakeyama, K. Shiren, K. Nakajima, S. Nomura, S. Nakatsuka, K. Kinoshita, J. Ni, Y. Ono, T. Ikuta. Ultrapure blue thermally activated delayed fluorescence molecules: Efficient homo–lumo separation by the multiple resonance effect. *Advanced Materials*, 28(14):2777–2781, 2016.

[45] K. Shizu, J. Lee, H. Tanaka, H. Nomura, T. Yasuda, H. Kaji, C. Adachi. Highly efficient electroluminescence from purely organic donor-acceptor systems. *Pure and Applied Chemistry*, 87(7):627–638, 2015.

[46] P. Rajamalli, N. Senthilkumar, P. Gandeepan, P. Y. Huang, M. J. Huang, C. Z. Ren-Wu, C. Y. Yang, M. J. Chiu, L. K. Chu, H. W. Lin, C. H. Cheng. A new molecular design based on thermally activated delayed fluorescence for highly efficient organic light emitting diodes. *Journal of the American Chemical Society*, 138(2):628–634, 2016.

[47] W. J. Park, Y. Lee, J. Y. Kim, D. W. Yoon, J. Kim, S. H. Chae, H. Kim, G. Lee, S. Shim, J. H. Yang, S. J. Lee. Effective thermally activated delayed fluorescence emitter and its performance in oled device. *Synthetic Metals*, 209(Supplement C):99–104, 2015.

[48] Y. J. Cho, K. S. Yook, J. Y. Lee. Cool and warm hybrid white organic light-emitting diode with blue delayed fluorescent emitter both as blue emitter and triplet host. *Scientific Reports*, 5(UNSP 7859), 2015.

[49] D. R. Lee, M. Kim, S. K. Jeon, S. H. Hwang, C. W. Lee, J. Y. Lee. Design strategy for 25% external quantum efficiency in green and blue thermally

activated delayed fluorescent devices. *Advanced Materials*, 27(39):5861–5867, 2015.

[50] M. Kim, S. K. Jeon, S.-H. Hwang, J. Y. Lee. Stable blue thermally activated delayed fluorescent organic light-emitting diodes with three times longer lifetime than phosphorescent organic light-emitting diodes. *Advanced Materials*, 27(15):2515–2520, 2015.

[51] Y. J. Cho, S. K. Jeon, B. D. Chin, E. Yu, J. Y. Lee. The design of dual emitting cores for green thermally activated delayed fluorescent materials. *Angewandte Chemie International Edition*, 54(17):5201–5204, 2015.

[52] Q. Zhang, B. Li, S. Huang, H. Nomura, H. Tanaka, C. Adachi. Efficient blue organic light-emitting diodes employing thermally activated delayed fluorescence. *Nature Photonics*, 8(4):326–332, 2014.

[53] Y. Kitamoto, T. Namikawa, D. Ikemizu, Y. Miyata, T. Suzuki, H. Kita, T. Sato, S. Oi. Light blue and green thermally activated delayed fluorescence from 10h-phenoxaborin-derivatives and their application to organic light-emitting diodes. *Journal of Materials Chemistry C*, 3(35):9122–9130, 2015.

[54] W.-L. Tsai, M.-H. Huang, W.-K. Lee, Y.-J. Hsu, K.-C. Pan, Y.-H. Huang, H.-C. Ting, M. Sarma, Y.-Y. Ho, H.-C. Hu, C.-C. Chen, M.-T. Lee, K.-T. Wong, C.-C. Wu. A versatile thermally activated delayed fluorescence emitter for both highly efficient doped and non-doped organic light emitting devices. *Chemical Communications*, 51:13662–13665, 2015.

[55] S. Wu, M. Aonuma, Q. Zhang, S. Huang, T. Nakagawa, K. Kuwabara, C. Adachi. High-efficiency deep-blue organic light-emitting diodes based on a thermally activated delayed fluorescence emitter. *Journal of Materials Chemestry C*, 2:421–424, 2014.

[56] C. Duan, J. Li, C. Han, D. Ding, H. Yang, Y. Wei, H. Xu. Multi-dipolar chromophores featuring phosphine oxide as joint acceptor: A new strategy

toward high-efficiency blue thermally activated delayed fluorescence dyes. *Chemistry of Materials*, 28(16):5667–5679, 2016.

[57] H. Ohkuma, T. Nakagawa, K. Shizu, T. Yasuda, C. Adachi. Thermally activated delayed fluorescence from a spiro-diazafluorene derivative. *Chemistry Letters*, 43(7):1017–1019, 2014.

[58] M. Liu, Y. Seino, D. Chen, S. Inomata, S. J. Su, H. Sasabe, J. Kido. Blue thermally activated delayed fluorescence materials based on bis(phenylsulfonyl)benzene derivatives. *Chemical Communications*, 51:16353–16356, 2015.

[59] J. W. Sun, J. Y. Baek, K.-H. Kim, C.-K. Moon, J.-H. Lee, S.-K. Kwon, Y.-H. Kim, J.-J. Kim. Thermally activated delayed fluorescence from azasiline based intramolecular charge-transfer emitter (dtpdda) and a highly efficient blue light emitting diode. *Chemistry of Materials*, 27(19):6675–6681, 2015.

[60] T. Takahashi, K. Shizu, T. Yasuda, K. Togashi, C. Adachi. Donor–acceptor-structured 1,4-diazatriphenylene derivatives exhibiting thermally activated delayed fluorescence: design and synthesis, photophysical properties and oled characteristics. *Science and Technology of Advanced Materials*, 15(3):034202, 2014.

[61] D. Y. Chen, W. Liu, C. J. Zheng, K. Wang, F. Li, S. L. Tao, X. M. Ou, X. H. Zhang. Isomeric thermally activated delayed fluorescence emitters for color purity-improved emission in organic light-emitting devices. *ACS Applied Materials & Interfaces*, 8(26):16791–16798, 2016.

[62] T.-A. Lin, T. Chatterjee, W.-L. Tsai, W.-K. Lee, M.-J. Wu, M. Jiao, K.-C. Pan, C.-L. Yi, C.-L. Chung, K.-T. Wong, C.-C. Wu. Sky-blue organic light emitting diode with 37% external quantum efficiency using thermally activated delayed fluorescence from spiroacridine-triazine hybrid. *Advanced Materials*, 28(32):6976–6983, 2016.

[63] Y. Liu, G. Xie, K. Wu, Z. Luo, T. Zhou, X. Zeng, J. Yu, S. Gong, C. Yang. Boosting reverse intersystem crossing by increasing donors in triarylboron/phenoxazine hybrids: Tadf emitters for high-performance solution-processed oleds. *Journal of Materials Chemistry C*, 4:4402–4407, 2016.

[64] G. Sliwoski, J. Mendenhall, J. Meiler. Autocorrelation descriptor improvements for qsar: 2da_sign and 3da_sign. *Journal of Computer-Aided Molecular Design*, 30(3):209–217, 2016.

[65] T. Puzyn. *Recent Advances in QSAR Studies: Methods and Applications*, volume 8 of *Challenges and Advances in Computational Chemistry and Physics*. Springer Netherlands, Dordrecht, 2010.

[66] R. Todeschini, V. Consonni. *Handbook of molecular descriptors*, volume 11. Wiley VCH, Weinheim, 01 2000.

[67] J. J. P. Stewart. Mopac2016, version: 17.240l.

[68] KRAKENX. `http://krakenminer.com/index.html`. Accessed: 10.01.2018.

[69] V. Venkatraman, B. K. Alsberg. Krakenx: software for the generation of alignment-independent 3d descriptors. *Journal of Molecular Modeling*, 22(4):93, 2016.

[70] N. Bodor, P. Buchwald, M.-J. Huang . Computer-assisted design of new drugs based on retrometabolic concepts. *SAR and QSAR in Environmental Research*, 8(1-2):41–92, 1998.

[71] N. Bodor, Z. Gabanyi, C. K. Wong,. A new method for the estimation of partition coefficient. *Journal of the American Chemical Society*, 111(11):3783–3786, 1989.

[72] A. Y. Meyer. The size of molecules. *Chem. Soc. Rev.*, 15:449–474, 1986.

[73] A. Y. Meyer. Molecular mechanics and molecular shape. v. on the computation of the bare surface area of molecules. *Journal of Computational Chemistry*, 9(1):18–24, 1988.

[74] G. A. Arteca. *Molecular Shape Descriptors*, chapter 5, pages 191–253. Wiley-Blackwell, 2007.

[75] D. T. Stanton, P. C. Jurs. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical Chemistry*, 62(21):2323–2329, 1990.

[76] A. O. Aptula, R. Kühne, R.-U. Ebert, M. T. D. Cronin, T. I. Netzeva, G. Schüürmann. Modeling discrimination between antibacterial and non-antibacterial activity based on 3d molecular descriptors. *QSAR & Combinatorial Science*, 22(1):113–128, 2003.

[77] G. Schüürmann. Qsar analysis of the acute fish toxicity of organic phosphorothionates using theoretically derived molecular descriptors. *Environmental Toxicology and Chemistry*, 9(4):417–428, 1990.

[78] K. Fukui. A new quantum-mechanical reactivity index for saturated compounds. *Bulletin of the Chemical Society of Japan*, 34(8), 1961.

[79] K. Fukui, T. Yonezawa, C. Nagata. Theory of substitution in conjugated molecules. *Bulletin of the Chemical Society of Japan*, 27(7):423–427, 1954.

[80] D. B. Turner, P. Willett. The eva spectral descriptor. *European Journal of Medicinal Chemistry*, 35(4):367 – 375, 2000.

[81] A. M. Ferguson, T. Heritage, P. Jonathon, S. E. Pack, L. Phillips, J. Rogan, P. J. Snaith. Eva: A new theoretically based molecular descriptor for use in qsar/qspr analysis. *Journal of Computer-Aided Molecular Design*, 11(2):143–152, 1997.

[82] K. Tuppurainen. Eeva (electronic eigenvalue): A new qsar/qspr descriptor for electronic substituent effects based on molecular orbital energies. *SAR and QSAR in Environmental Research*, 10(1):39–46, 1999.

[83] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[84] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36, 2014.

[85] A. Kassambara, F. Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2017. R package version 1.0.5.

[86] B.-H. Mevik, R. Wehrens, K. H. Liland. *pls: Partial Least Squares and Principal Component Regression*, 2016. R package version 2.6-0.

[87] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[88] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011.

[89] A. Day. Contributions from T. L. Davis and R. Zenka. *getopt: C-Like 'getopt' Behavior*, 2017. R package version 1.20.1.

[90] M. Kuhn, R. Quinlan. *Cubist: Rule- And Instance-Based Regression Modeling*, 2017. R package version 0.2.1.

[91] M. Kuhn. Contributions from J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, T. Hunt. *caret: Classification and Regression Training*, 2018. R package version 6.0-79.

[92] Microsoft Corporation and S. Weston. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, 2017. R package version 1.0.11.

[93] M. Valiev, E.J. Bylaska, N. Govind, K. Kowalski, T.P. Straatsma, H.J.J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T.L. Windus, W.A. de Jong. Nwchem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Computer Physics Communications*, 181(9):1477 – 1489, 2010.

[94] A. Klamt, G. Schüürmann, G. Cosmo: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, 2(5):799–805, 1993.

[95] D. M. York, M. Karplus. A smooth solvation potential based on the conductor-like screening model. *The Journal of Physical Chemistry A*, 103(50):11060–11079, 1999.

[96] T. Næs, H. Martens. *Multivariate Calibration*. John Wiley & Sons, 1991.

[97] B. K. Alsberg. Chemometrics. Compendium.

[98] S. Wold, K. Esbensen, P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37 – 52, 1987.

[99] H. Martens. *Multivariate analysis of quality : an introduction*. Wiley, Chichester, 2001.

[100] P. Geladi, B. R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1 – 17, 1986.

[101] S. Wold, M. Sjöström, L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109 – 130, 2001.

[102] M. Ruiz, L. E. Mujica, X. Berjaga, J. Rodellar. Partial least square/projection to latent structures (pls) regression to estimate impact localization in structures. *Smart Materials and Structures*, 22, 2013.

[103] Y. Dodge. *Randomization*, pages 447–447. Springer New York, New York, NY, 2008.

[104] S. Wiklund, D. Nilsson, L. Eriksson, M. Sjöström, S. Wold, K. Faber. A randomization test for pls component selection. *Journal of Chemometrics*, 21(10-11):427–439, 2007.

[105] C. M. Andersen, R. Bro. Variable selection in regression—a tutorial. *Journal of Chemometrics*, 24(11-12):728–737, 2010.

[106] J. Willmott, C and Matsuura, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30:79, 2005.

[107] G. Schüürmann, R. -U. Ebert, J. Chen, B. Wang, R. Kühne. External validation and prediction employing the predictive squared correlation coefficient — test set activity mean vs training set activity mean. *Journal of Chemical Information and Modeling*, 48(11):2140–2145, 2008.

[108] R. Todeschini. *Tools for prediction of environmental properties of chemiclas by QSAR/QSPR within reach. An applicability domain perspective.* PhD thesis, Scuola di dottorato di Scienze, Italy, 2013.

[109] Q. Zang, K. Mansouri, A. J. Williams, R. S. Judson, D. G. Allen, W. M. Casey, N. C. Kleinstreuer. In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *Journal of Chemical Information and Modeling*, 57(1):36–49, 2017.

[110] M. H. Kutner. *Applied Linear Statistical Models.* McGrwa-Hill international edition. McGraw-Hill Irwin, 2005.

[111] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[112] Y. Dodge. *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY, 2008.

[113] J. R. Quinlan. Learning with continuous classes. In *Proceedings of the 5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348, 1992.

[114] J. R. Quinlan. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on International onference on*

Machine Learning, pages 236–243, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[115] J. R. Quinlan. Simplifying decision trees. *International Journal of Human - Computer Studies*, 51(2):497–510, 1999.

[116] M. Kuhn, K. Johnson. *Applied Predictive Modeling*. Springer New York, New York, NY, 2013.

[117] T. J. Hastie. The elements of statistical learning : data mining, inference, and prediction, 2001.

[118] F. Jensen. *Introduction to computational chemistry*. Wiley, Chichester, 2nd ed. edition, 2007.

[119] P. Hohenberg, W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, 1964.

[120] W. Koch, M. C. Holthausen. *A Chemist's Guide to Density Functional Theory*. WILEY-VCH, 2 edition, 2001.

[121] W. Kohn, L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965.

[122] Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review. A, General physics*, 38(6), 1988.

[123] G. J. Laming, V. Termath, N. C. Handy. A general purpose exchange-correlation energy functional. *The Journal of Chemical Physics*, 99(11):8765–8773, 1993.

[124] C. Lee, W. Yang, R. G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, 1988.

[125] R. Colle, O. Salvetti. Approximate calculation of the correlation energy for the closed shells. *Theoretica chimica acta*, 37(4):329–334, 1975.

[126] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch. Ab initio calculations of vibrational absorption and circular-dichromism spectra using density functional force fields. *Journal Of Physical Chemistry*, 98(45):11623–11627, 1994.

[127] M. E. Casida. Time-dependent density functional response theory for molecules. In *Recent Advances In Density Functional Methods*, pages 155–192. World Scientific Publishing Co. Pte. Ltd., 1995.

[128] R. Bauernschmitt, R. Ahlrichs. Treatment of electronic excitations within the adiabatic approximation of time dependent density functional theory. *Chemical Physics Letters*, 256(4):454 – 464, 1996.

[129] D. J. Tozer, N. C. Handy. Improving virtual kohn–sham orbitals and eigenvalues: Application to excitation energies and static polarizabilities. *The Journal of Chemical Physics*, 109(23):10180–10189, 1998.

[130] D. J. Tozer, R. D. Amos, N. C. Handy, B. O. Roos, L. Serrano-Andres. Does density functional theory contribute to the understanding of excited states of unsaturated organic compounds? *Molecular Physics*, 97(7):859–868, 1999.

[131] A. Dreuw, J. L. Weisman, M. Head-Gordon. Long-range charge-transfer excited states in time-dependent density functional theory require non-local exchange. *The Journal of Chemical Physics*, 119(6):2943–2946, 2003.

[132] L. Bernasconi, M. Sprik, J. Hutter. Time dependent density functional theory study of charge-transfer and intramolecular electronic excitations in acetone–water systems. *The Journal of Chemical Physics*, 119(23):12417–12431, 2003.

[133] T. Yanai, D. P. Tew, N. C. Handy. A new hybrid exchange–correlation functional using the coulomb-attenuating method (cam-b3lyp). *Chemical Physics Letters*, 393(1):51–57, 2004.

[134] M. Swart. A new family of hybrid density functionals. *Chemical Physics Letters*, 580:166–171, 2013.

[135] A. D. Becke. A multicenter numerical integration scheme for polyatomic molecules. *The Journal of Chemical Physics*, 88(4):2547–2553, 1988.

[136] C. W. Murray, N. C. Handy, G. J. Laming. Quadrature schemes for integrals of density functional theory. *Molecular Physics*, 78(4):997–1014, 1993.

[137] M. E. Mura, P. J. Knowles. Improved radial grids for quadrature in molecular density-functional calculations. *The Journal of Chemical Physics*, 104(24):9848–9858, 1996.

[138] V. I.Lebedev. Values of the nodes and weights of ninth to seventeenth order gauss-markov quadrature formulae invariant under the octahedron group with inversion. *USSR Computational Mathematics and Mathematical Physics*, 15(1):44–51, 1975.

[139] V. I. Lebedev. Quadratures on a sphere. *USSR Computational Mathematics and Mathematical Physics*, 16(2):10–24, 1976.

[140] V. I. Lebedev. Spherical quadrature formulas exact to orders 25–29. *Siberian Mathematical Journal*, 18(1):99–107, 1977.

[141] V. I. Lebedev, A. L. Skorokhodov. Quadrature formulas of orders 41, 47, and 53 for the sphere. *Russian Academy of Sciences Doklady Mathematics*, 45:587–592, 1992.

[142] V. I. Lebedev. A quadrature formula for the sphere of 59th algebraic order of accuracy. *Russian Academy of Sciences Doklady Mathematics*, 50:283–286, 1995.

[143] V. I. Lebedev. A quadrature formula for the sphere of the 131-st algebraic order of accuracy. *Russian Academy of Sciences Doklady Mathematics*, 59:477–481, 1999.

[144] F. Jensen. Polarization consistent basis sets: Principles. *The Journal of Chemical Physics*, 115(20):9113–9125, 2001.

[145] R. C. Raffenetti. General contraction of gaussian atomic orbitals: Core, valence, polarization, and diffuse basis sets; molecular integral evaluation. *The Journal of Chemical Physics*, 58(10):4452–4458, 1973.

[146] T. H. Dunning. Gaussian basis functions for use in molecular calculations. i. contraction of (9s5p) atomic basis sets for the first-row atoms. *The Journal of Chemical Physics*, 53(7):2823–2833, 1970.

[147] F. Jensen. Unifying general and segmented contracted basis sets. segmented polarization consistent basis sets. *Journal of Chemical Theory and Computation*, 10(3):1074–1085, 2014.

[148] E. R. Davidson. Comment on "comment on dunning's correlation-consistent basis sets". *Chemical Physics Letters*, 260(3):514 – 518, 1996.

[149] F. Jensen. Erratum: "polarization consistent basis sets: Principles" [j. chem. phys. 115, 9113 (2001)]. *The Journal of Chemical Physics*, 116(8):3502–3502, 2002.

[150] F. Jensen, T. Helgaker. Polarization consistent basis sets. v. the elements si–cl. *The Journal of Chemical Physics*, 121(8):3463–3470, 2004.

[151] F. Jensen. Polarization consistent basis sets. 4: The elements he, li, be, b, ne, na, mg, al, and ar. *The Journal of Physical Chemistry A*, 111(44):11198–11204, 2007.

[152] F. Jensen. Polarization consistent basis sets. vii. the elements k, ca, ga, ge, as, se, br, and kr. *The Journal of Chemical Physics*, 136(11):114107, 2012.

[153] F. Jensen. Polarization consistent basis sets. viii. the transition metals sc-zn. *The Journal of Chemical Physics*, 138(1):014107, 2013.

[154] E. Runge, E. K. U. Gross. Density-functional theory for time-dependent systems. *Phys. Rev. Lett.*, 52:997–1000, 1984.

[155] R. van Leeuwen. Mapping from densities to potentials in time-dependent density-functional theory. *Phys. Rev. Lett.*, 82:3863–3866, 1999.

[156] A. Castro, H. Appel, M. Oliveira, C. A. Rozzi, X. Andrade, F. Lorenzen, M. A. L. Marques, E. K. U. Gross, A. Rubio. octopus: a tool for the application of time-dependent density functional theory. *physica status solidi (b)*, 243(11):2465–2488, 2006.

[157] M. Petersilka, U. J. Gossmann, E. K. U. Gross. Excitation energies from time-dependent density-functional theory. *Phys. Rev. Lett.*, 76:1212–1215, 1996.

[158] M. A. L Marques. *Time-Dependent Density Functional Theory*, volume 706 of *Lecture Notes in Physics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[159] E. K. U. Gross, N. T. Maitra. Introduction to tddft. In F. M. S. Nogueira E. K. U. Gross A. Rubio M. A. L. Marques, N. T. Maitra, editor, *Fundamentals of Time-Dependent Density Functional Theory*, pages 53–99, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[160] G. D. Temple. *Time-Dependent Density Functional Theory for Open Quantum Systems and Quantum Computation*. PhD thesis, Harvard University, 2012.

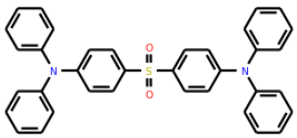
[161] N. T. Maitra. Memory formulas for perturbations in time-dependent density functional theory. *International Journal of Quantum Chemistry*, 102(5):573–581, 2005.

[162] N. T. Maitra, K. Burke. Demonstration of initial-state dependence in time-dependent density-functional theory. *Phys. Rev. A*, 63:042501, 2001.

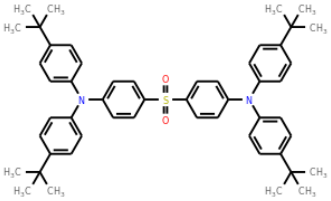
[163] N. T. Maitra, K. Burke, C. Woodward. Memory in time-dependent density functional theory. *Phys. Rev. Lett.*, 89:023002, 2002.

[164] K. A. Dill. Molecular driving forces : statistical thermodynamics in biology, chemistry, physics, and nanoscience, 2011.

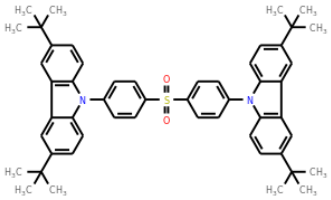
A 2D representation of all the structures



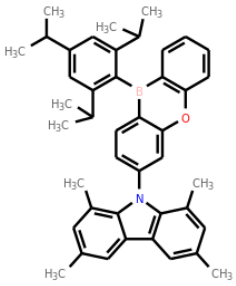
(i) 1 [26]



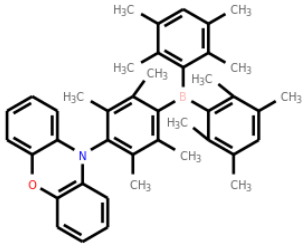
(ii) 2 [26]



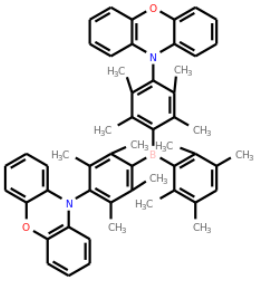
(iii) 3 [26]



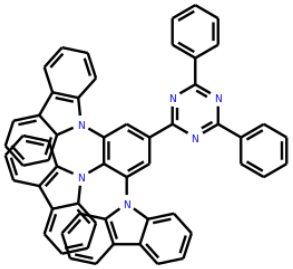
(iv) TMCPOB [34]



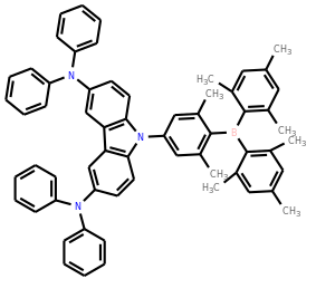
(v) TB-1PXZ [63]



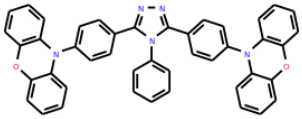
(vi) TB-2PXZ [63]



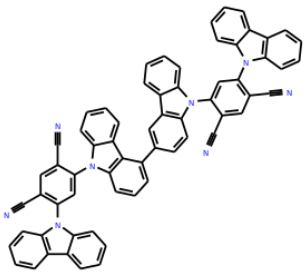
(vii) TCzTrz [49]



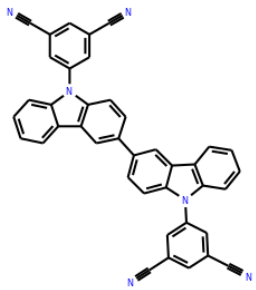
(viii) 2DAC-Mes3B [27]



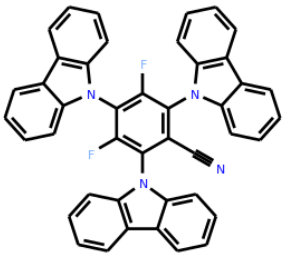
(ix) 2PXZ-TAZ [28]



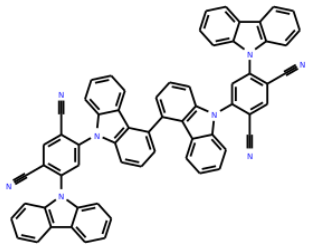
(x) 34TCzPN [29]



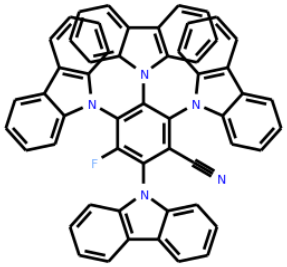
(xi) 35IPNDCz [30]



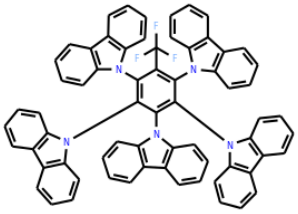
(xii) 3CzFCN [31]



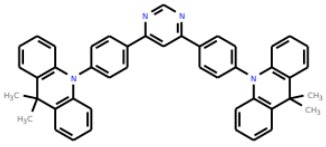
(xiii) 44TCzPN [29]



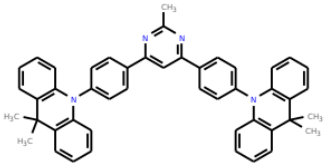
(xiv) 4CzFCN [31]



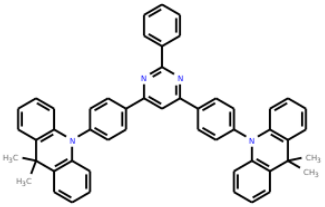
(xv) 5CzCF3Ph [32]



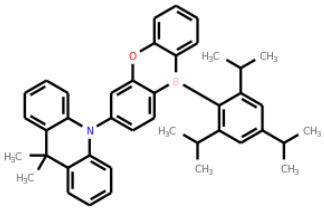
(xvi) Ac-HPM [33]



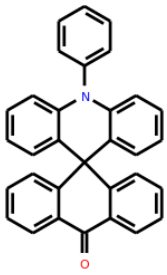
(xvii) Ac-MPM [33]



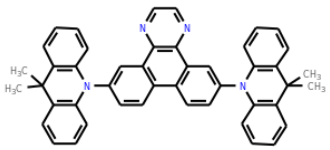
(xviii) Ac-PPM [33]



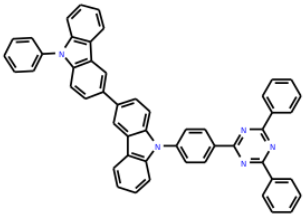
(xix) ACRPOB [34]



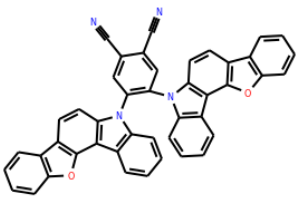
(xx) ACRSA [35]



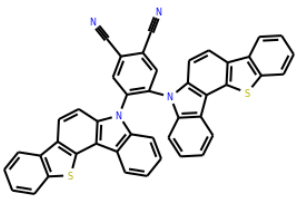
(xxi) ATP-ACR [36]



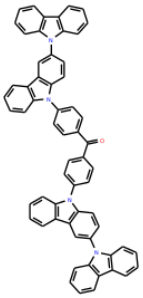
(xxii) BCzT [37]



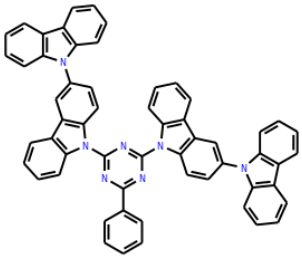
(xxiii) BFCz-2CN [38]



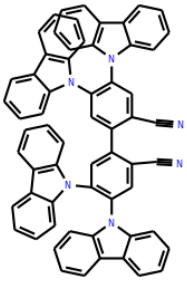
(xxiv) BTCz-2CN [38]



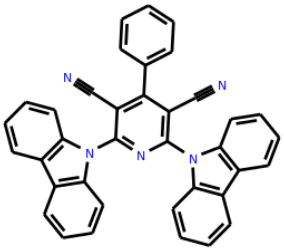
(xxv) CC2BP [39]



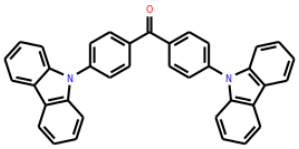
(xxvi) CC2TA [40]



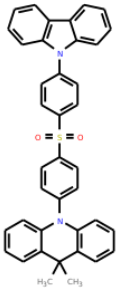
(xxvii) CNBPCz [41]



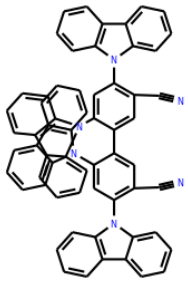
(xxviii) CPC [42]



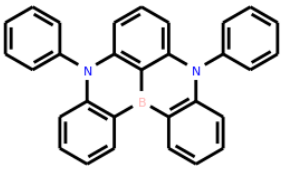
(xxix) Cz2BP [39]



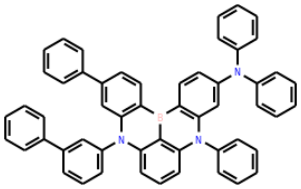
(xxx) CzAcSF [43]



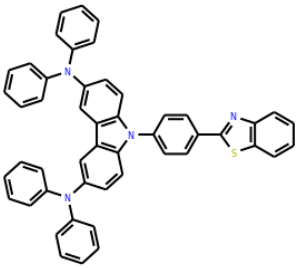
(xxxi) CzBPCN [41]



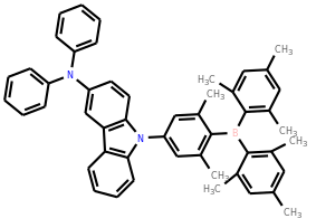
(xxxii) DABNA-1 [44]



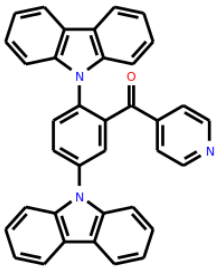
(xxxiii) DABNA-2 [44]



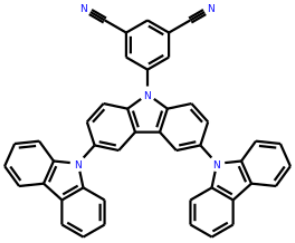
(xxxiv) DAC-BTZ [45]



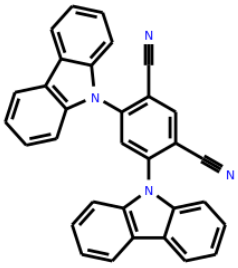
(xxxv) DAC-Mes3B [27]



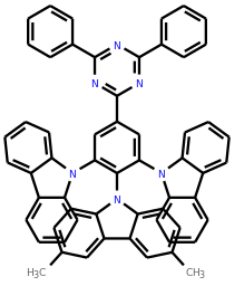
(xxxvi) DCBPy [46]



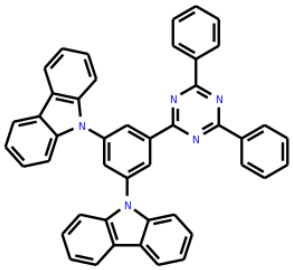
(xxxvii) DCN-3 [47]



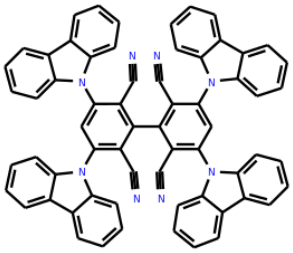
(xxxviii) DCzIPN [48]



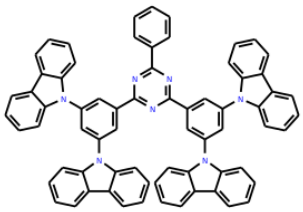
(xxxix) DCzmCzTrz [49]



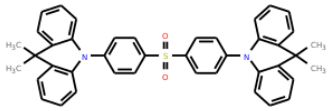
(xl) DCzTrz [50]



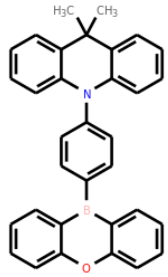
(xli) DDCzIPN [51]



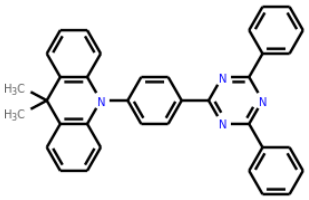
(xlii) DDCzTrz [50]



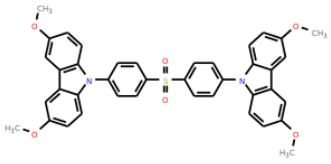
(xliii) DMAC-DPS [52]



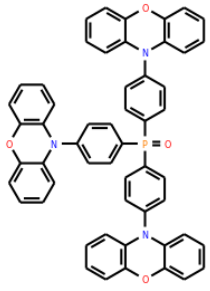
(xliv) DMAC-PXB [53]



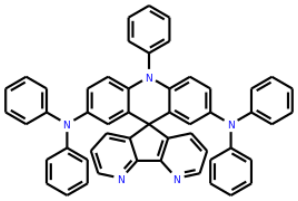
(xlv) DMAC-TRZ [54]



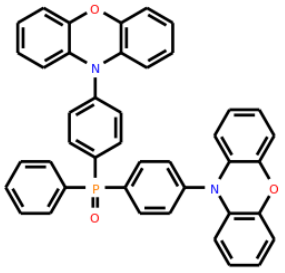
(xlvi) DMOC-DPS [55]



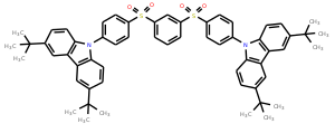
(xlvii) TPXZPO [56]



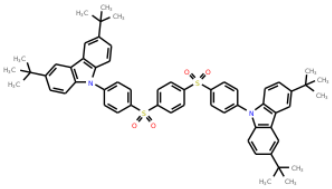
(xlviii) DPAA-AF [57]



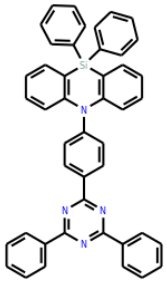
(xlix) DPXZPO [56]



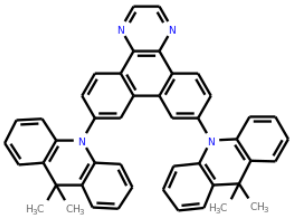
(l) DTC-mBPSB [58]



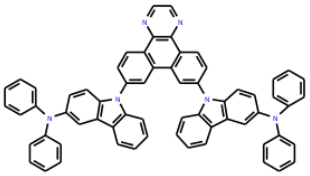
(li) DTC-pBPSB [58]



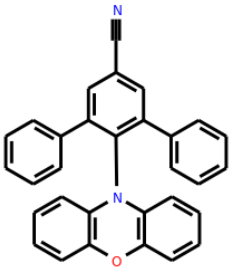
(lii) DTPDDA [59]



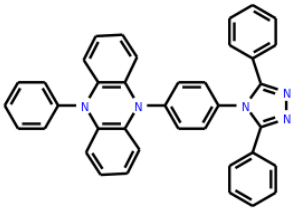
(liii) m-ATP-ACR [60]



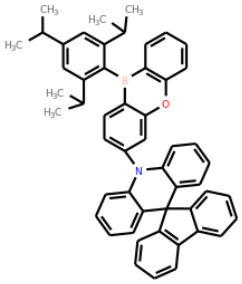
(liv) m-ATP-CDP [60]



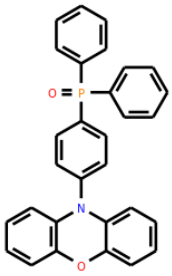
(lv) mPTC [61]



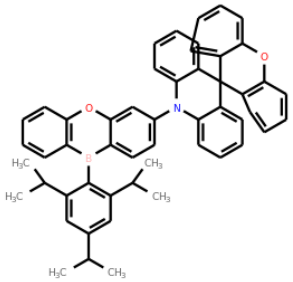
(lvi) PPZ-4TPT [52]



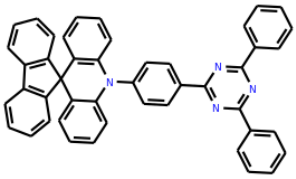
(lvii) SFDPAPOB [34]



(lviii) SPXZPO [56]



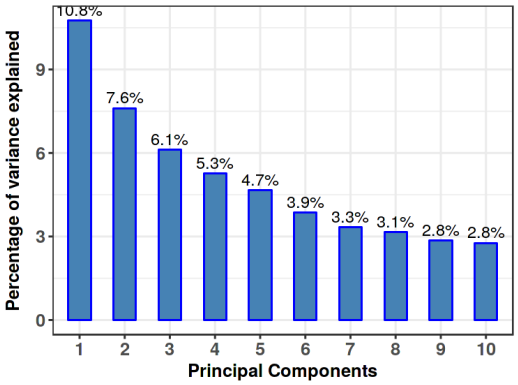
(lix) SXDPAPOB [34]



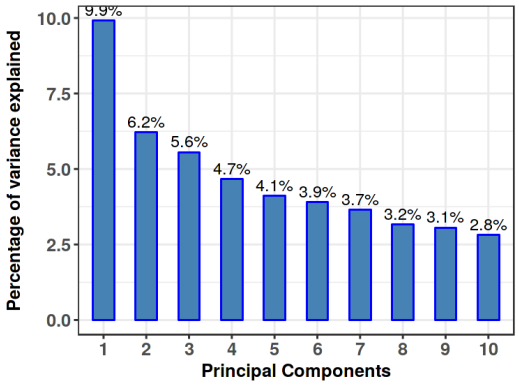
(lx) SpiroAC-TRZ [62]

B PCA result

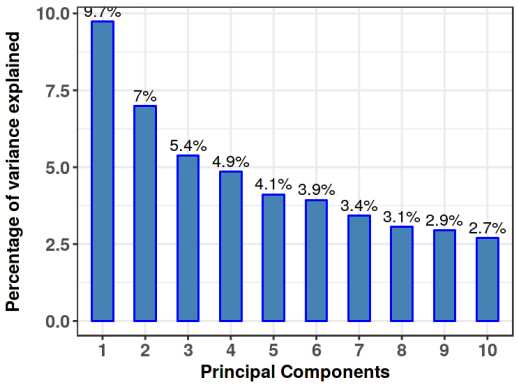
B.1 Full data set



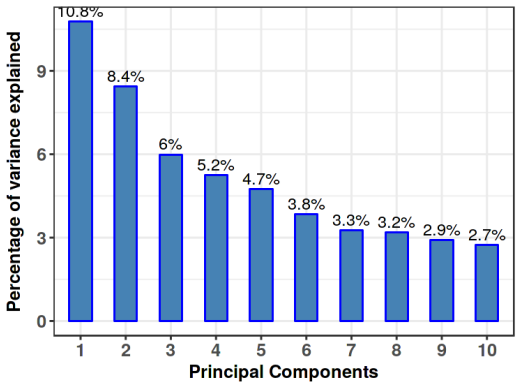
(i) EVA



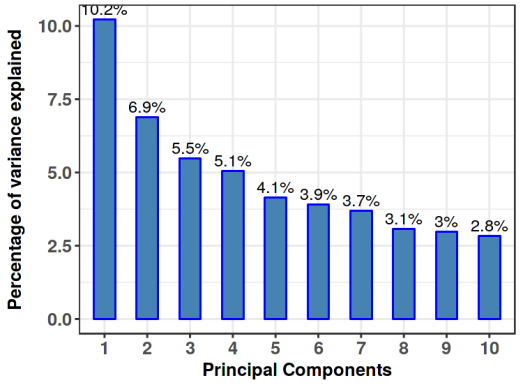
(ii) EEVA



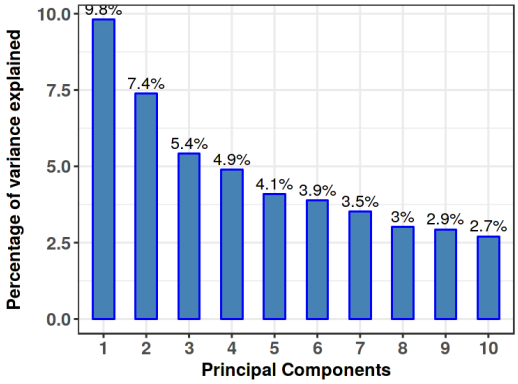
(iii) EVA and EEVA



(iv) EVA and REST



(v) EEVA and REST



(vi) ALL

Figure B.1: Variance explained by the first ten PCs for the PCA with different descriptor sets. Obtained during the master project, fall 2017.

B.2 Identification of potential outliers

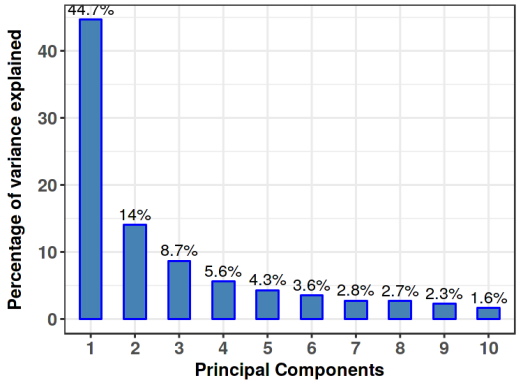
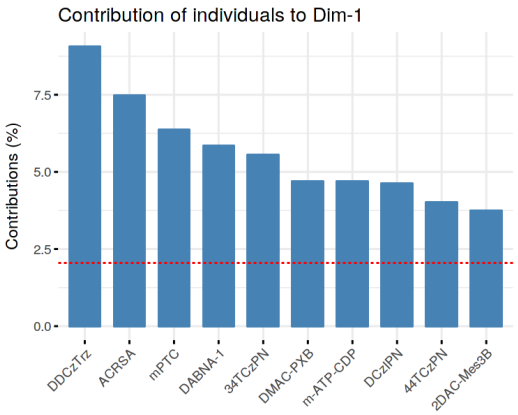
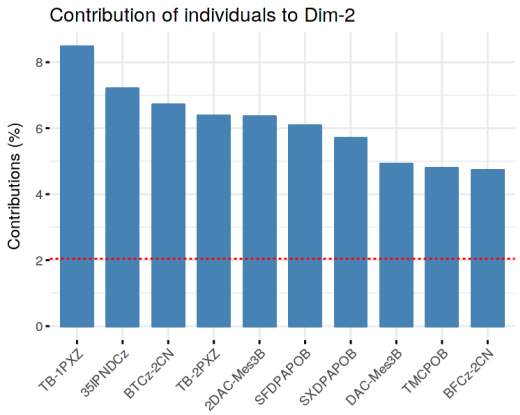


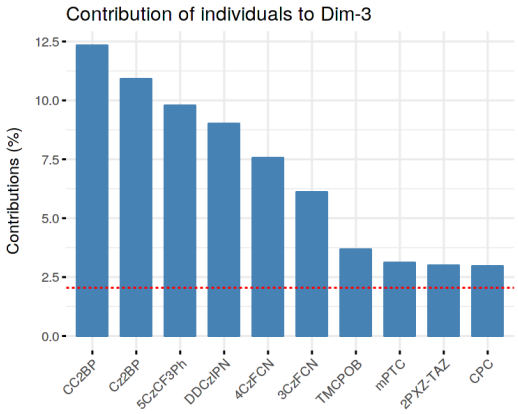
Figure B.2: Variance explained by the first ten PCs for the case where the eleven possible outliers are removed and the REST set of descriptors is used.



(i) Object contribution to PC 1

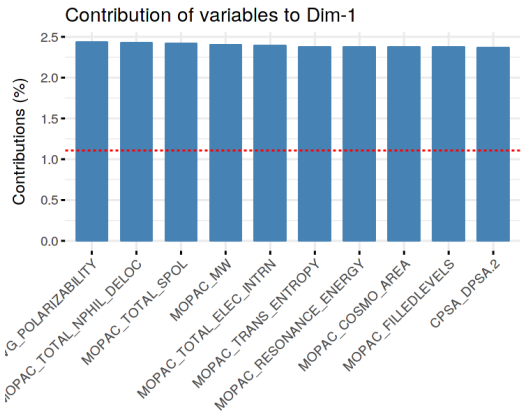


(ii) Object contribution to PC 2

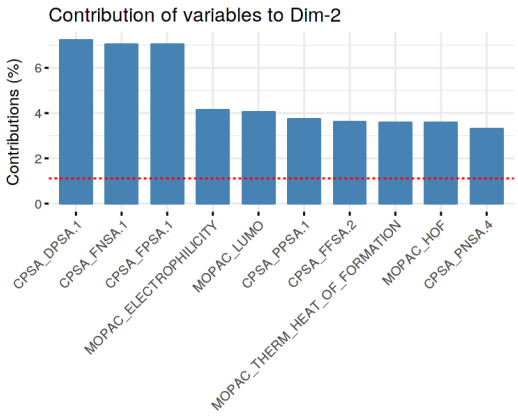


(iii) Object contribution to PC 3

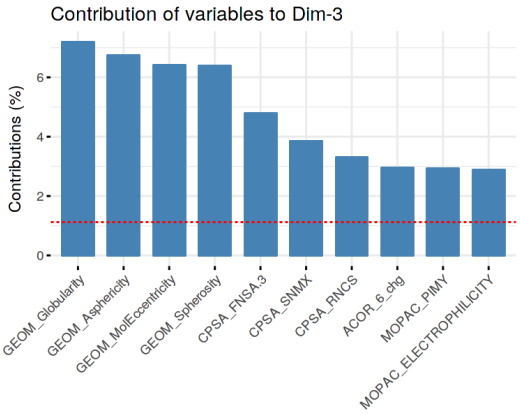
Figure B.3: Object contribution to the first three PCs for data set where the eleven potential outliers are removed. The REST set of descriptors is used.



(i) Variable contribution to PC 1



(ii) Variable contribution to PC 2



(iii) Variable contribution to PC 3

Figure B.4: Variable contribution to the first three PCs for the data set where the eleven potential outliers are removed. The REST set of descriptors is used.

B.3 Solvent effects

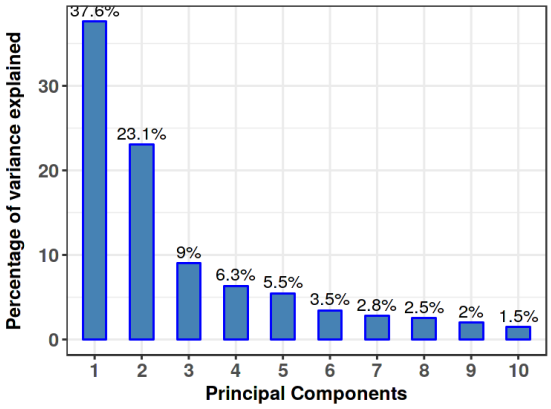
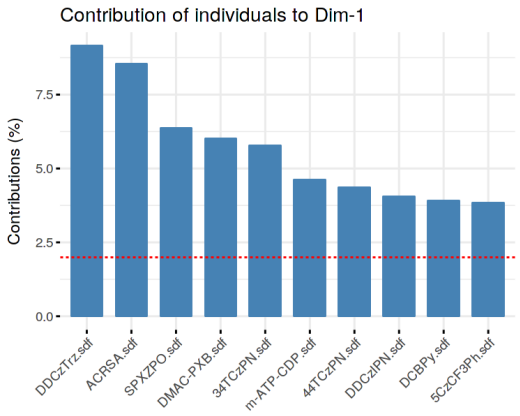
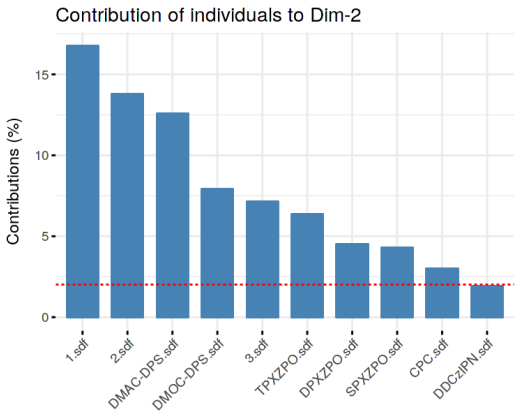


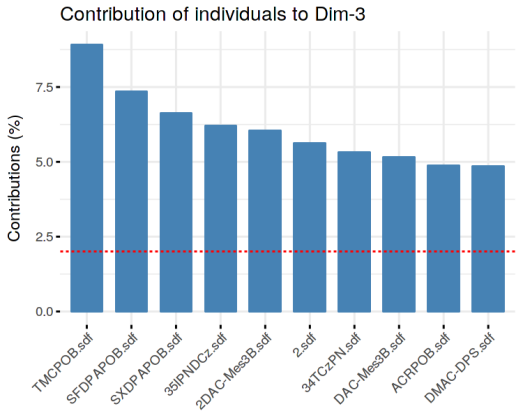
Figure B.5: Variance explained by the first ten PCs for when the REST set of descriptors is used and only the objects with experimental ΔE_{st} values measured in toluene are considered.



(i) Object contribution to PC 1

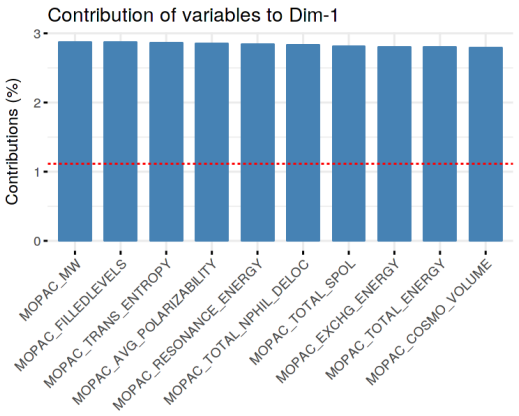


(ii) Object contribution to PC 2

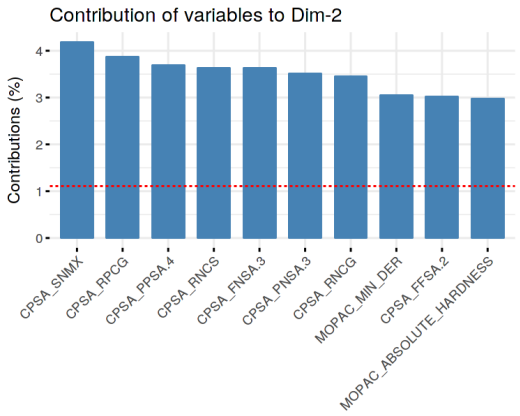


(iii) Object contribution to PC 3

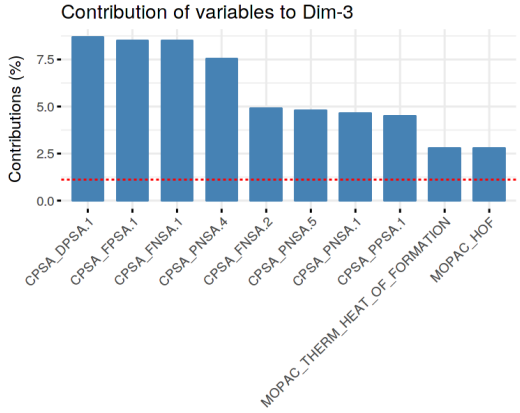
Figure B.6: Object contribution to the first three PCs when the REST set of descriptors is used and only the objects with experimental ΔE_{st} values measured in toluene are considered.



(i) Variable contribution to PC 1



(ii) Variable contribution to PC 2



(iii) Variable contribution to PC 3

Figure B.7: Variable contribution to the first three PCs for when the REST set of descriptors is used and only the objects with experimental ΔE_{st} values measured in toluene are considered.

C PLSR result

C.1 Full data set

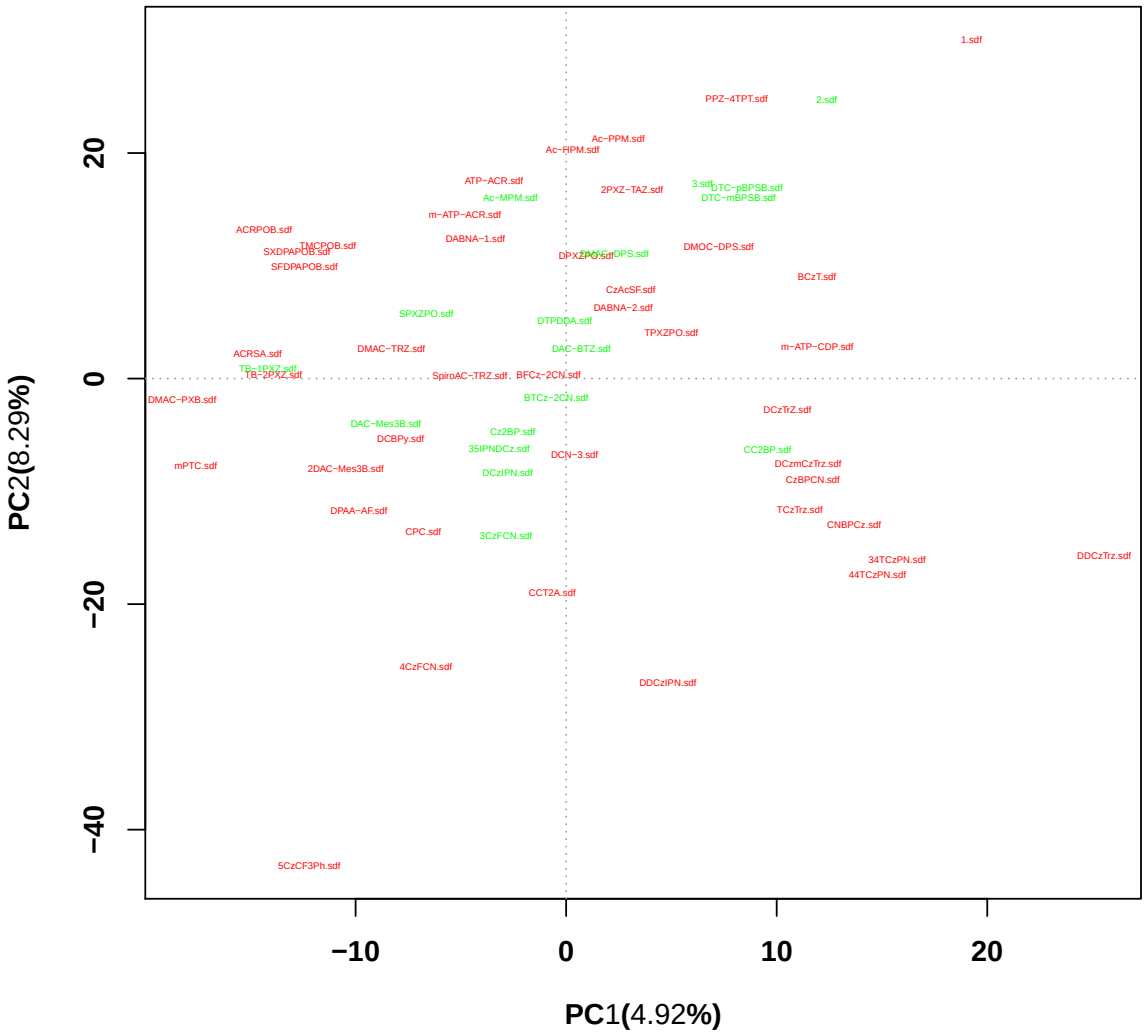


Figure C.1: Scores plot full data set with all descriptors included and no variable selection is performed. PC 2 plotted against PC 1, where the percentage represents the X-variance.

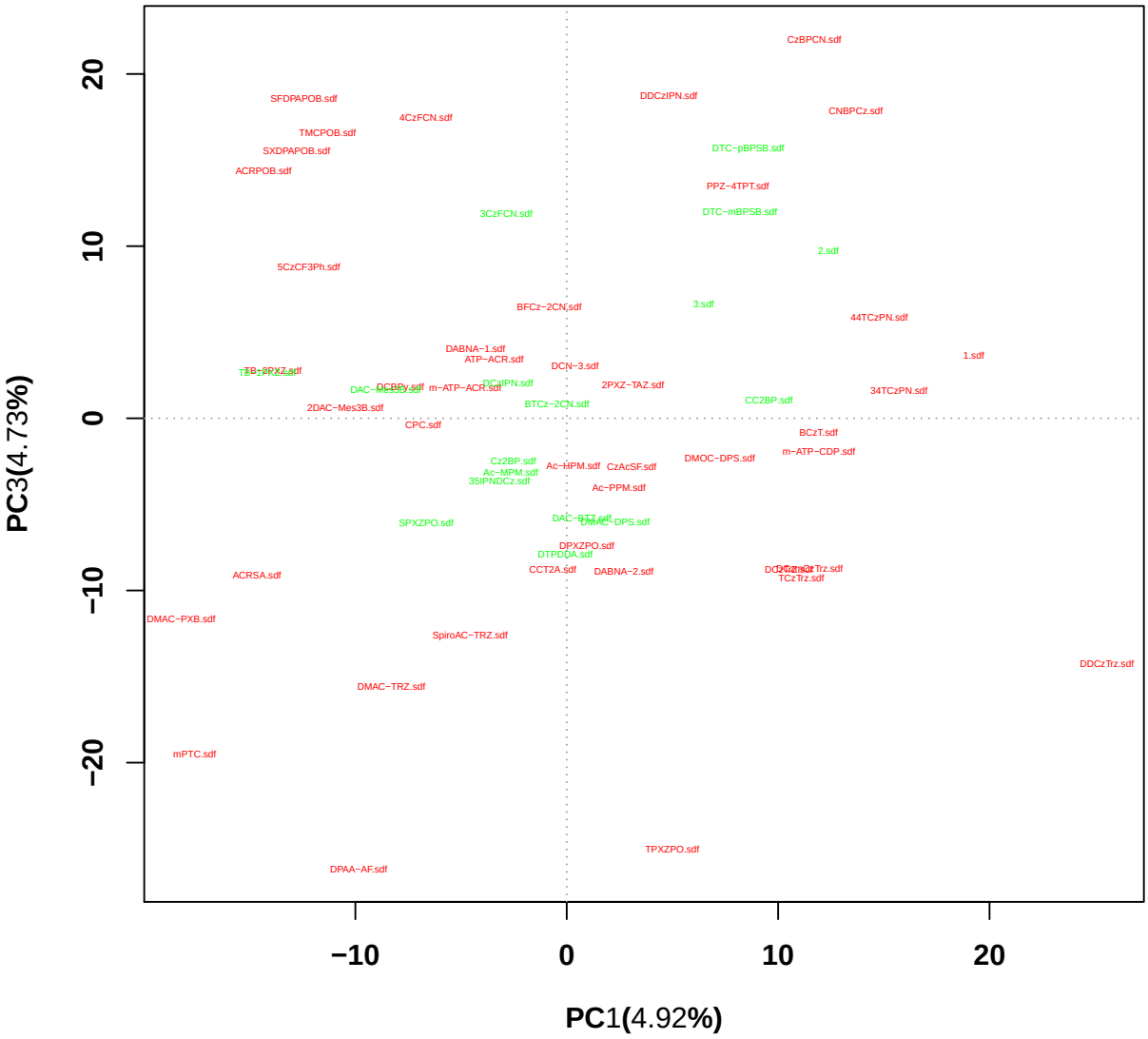


Figure C.2: Scores plot full data set with all descriptors included and no variable selection is performed. PC 3 plotted against PC 1, where the percentage represents the X-variance.

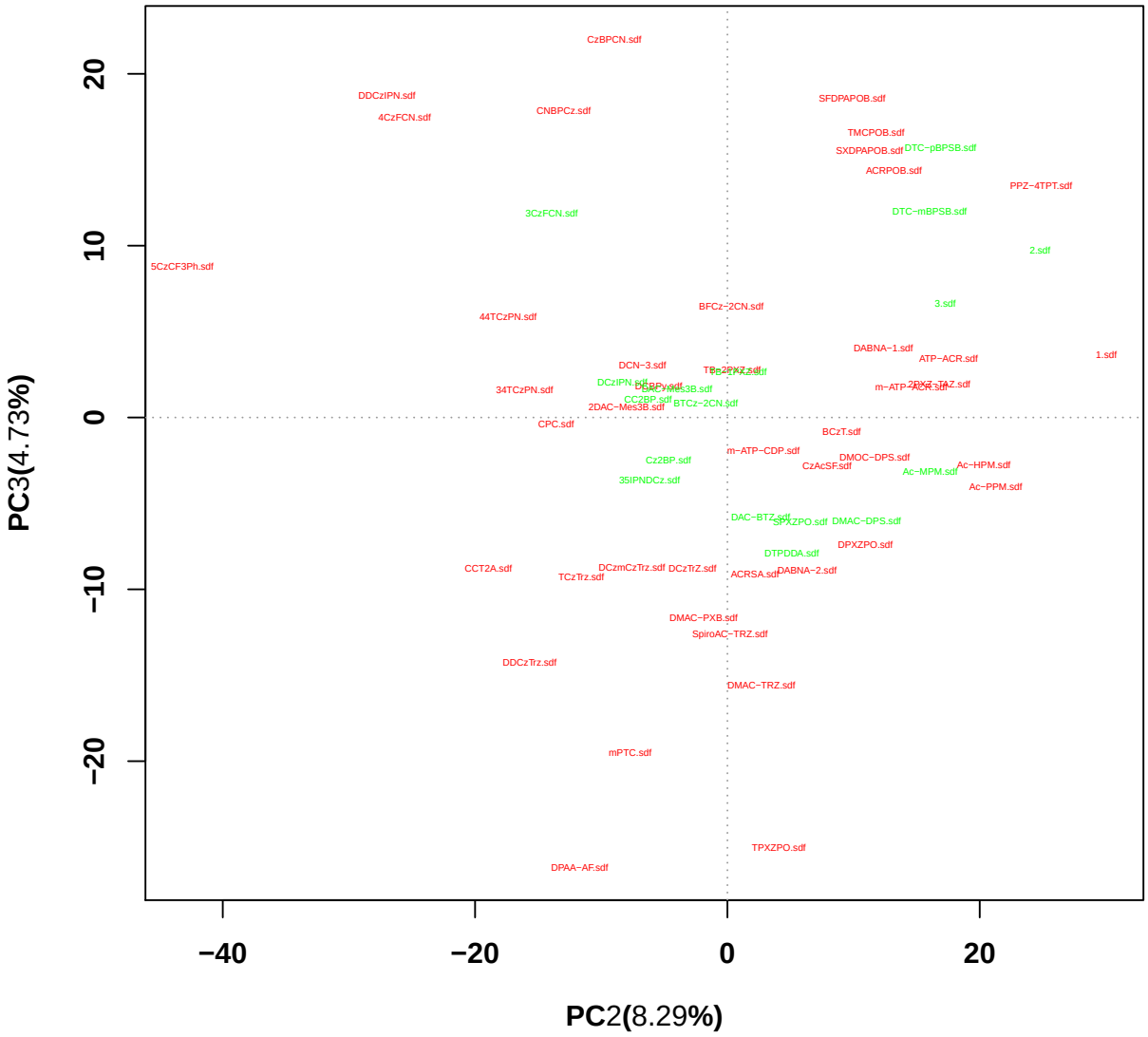


Figure C.3: Scores plot full data set with all descriptors included and no variable selection is performed. PC 3 plotted against PC 2, where the percentage represents the X-variance.

C.2 Identification of potential outliers

Table C.1: Observed and predicted ΔE_{st} values for the test set, ΔE_{st}^{obs} and ΔE_{st}^{pred}, respectively. The eleven potential outliers are excluded. All descriptors are included and no variable selection is performed. Note that this is the third root of the experimental ΔE_{st} values.

Structure	ΔE_{st}^{obs}	ΔE_{st}^{pred}
35IPNDCz	0.52	0.46
3CzFCN	0.39	0.46
Ac-MPM	0.58	0.50
CNBPCz	0.65	0.52
CPC	0.34	0.45
Cz2BP	0.59	0.43
CzBPCN	0.65	0.53
DABNA-1	0.57	0.47
DCzIPN	0.37	0.43
m-ATP-ACR	0.51	0.48
SFDPAPOB	0.45	0.41
SXDPAPOB	0.45	0.43

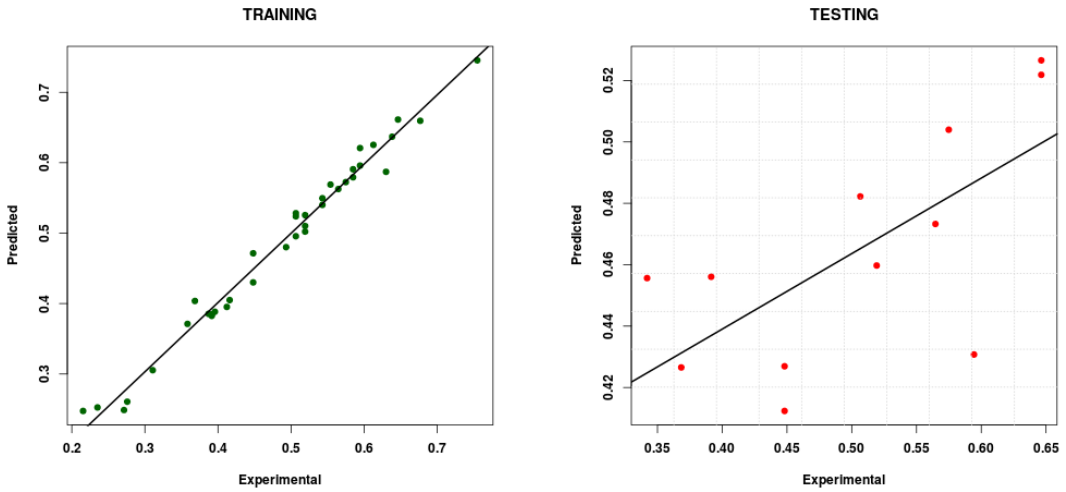


Figure C.4: Fit of the predicted values in the training set and the testing set when all the descriptors are included. The eleven potential outliers are excluded. The y- and x-axis represents the predicted and the experimental ΔE_{st} values, respectively.

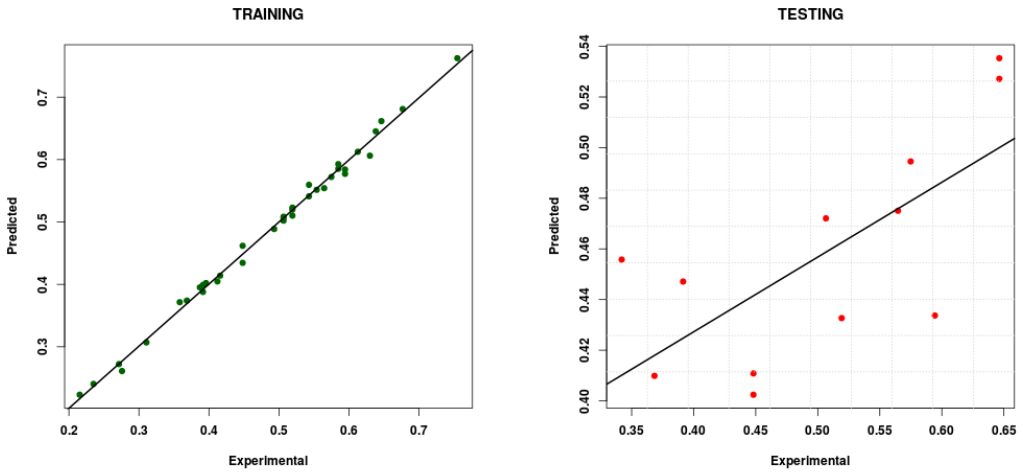


Figure C.5: Fit of the predicted values in the training set and the testing set when all the descriptors are included and variable selection is performed with VIP=0,8. The eleven potential outliers are excluded. The y- and x-axis represents the predicted and the experimental ΔE_{st} values, respectively.

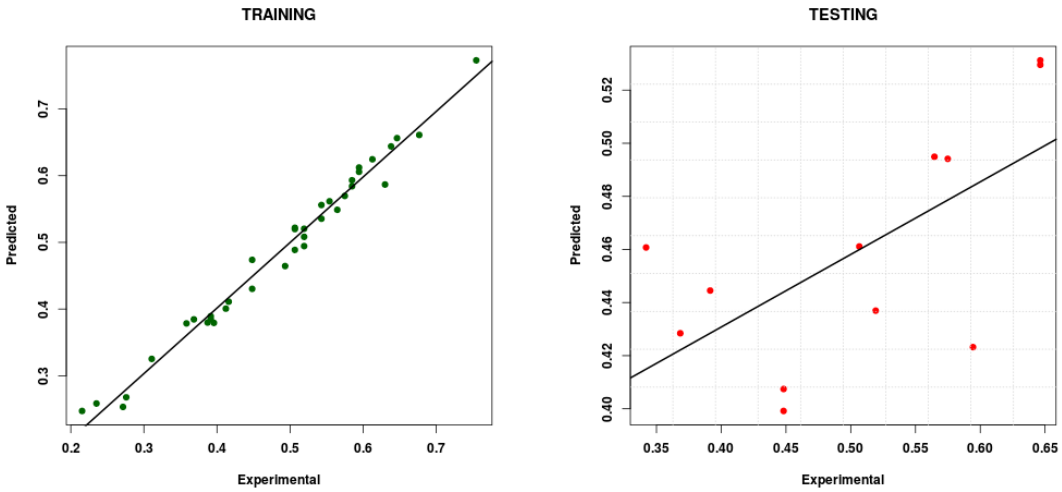


Figure C.6: Fit of the predicted values in the training set and the testing set when all the descriptors are included and variable selection is performed with VIP=1. The eleven potential outliers are excluded. The y- and x-axis represents the predicted and the experimental ΔE_{st} values, respectively.

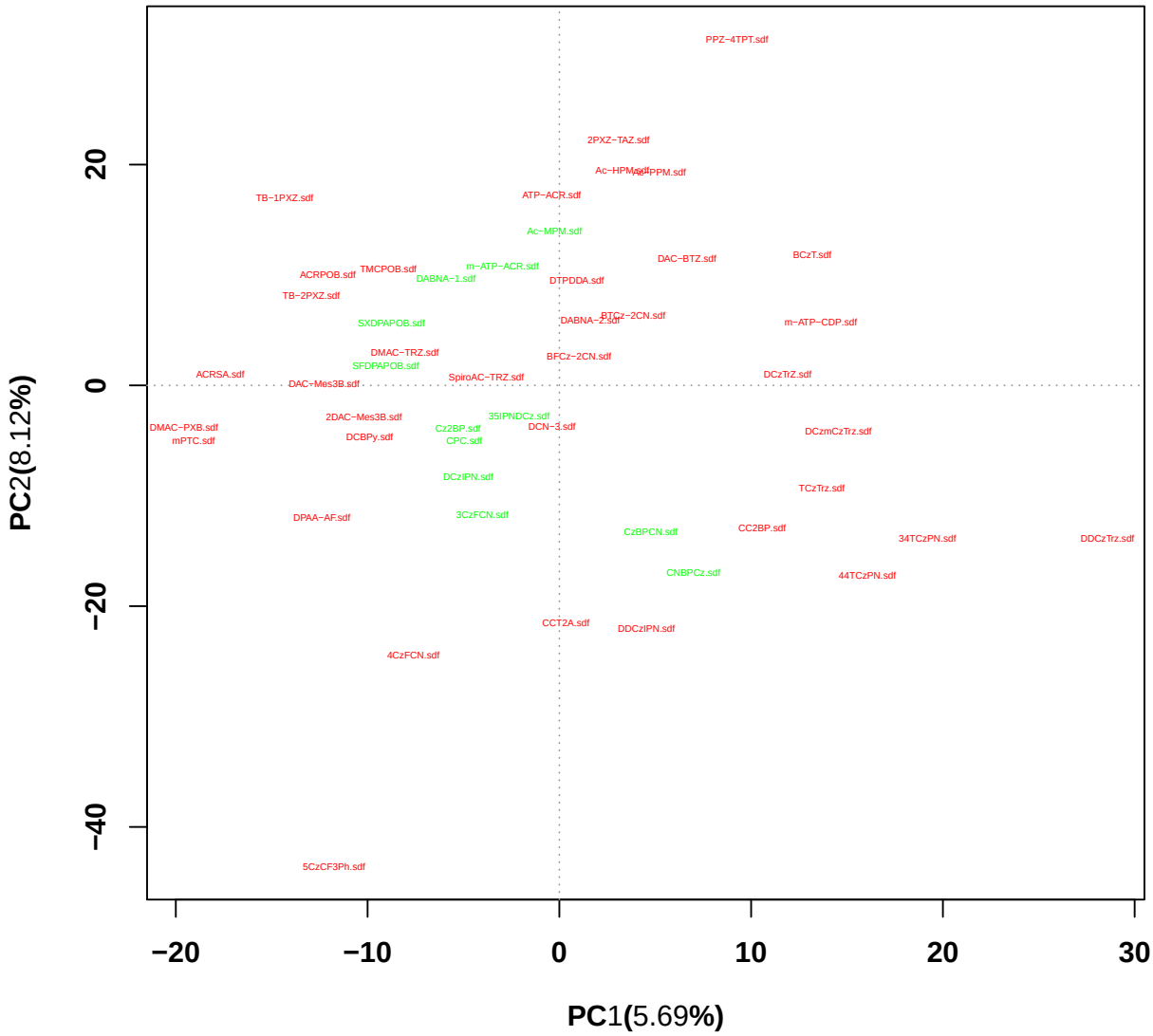


Figure C.7: Scores plot from the PLSR when the eleven potential outliers are removed. All the descriptors are included and no variable selection is performed. PC 2 is plotted against PC 1 and the descriptor variance explained by each of these PCs is given in parenthesis.

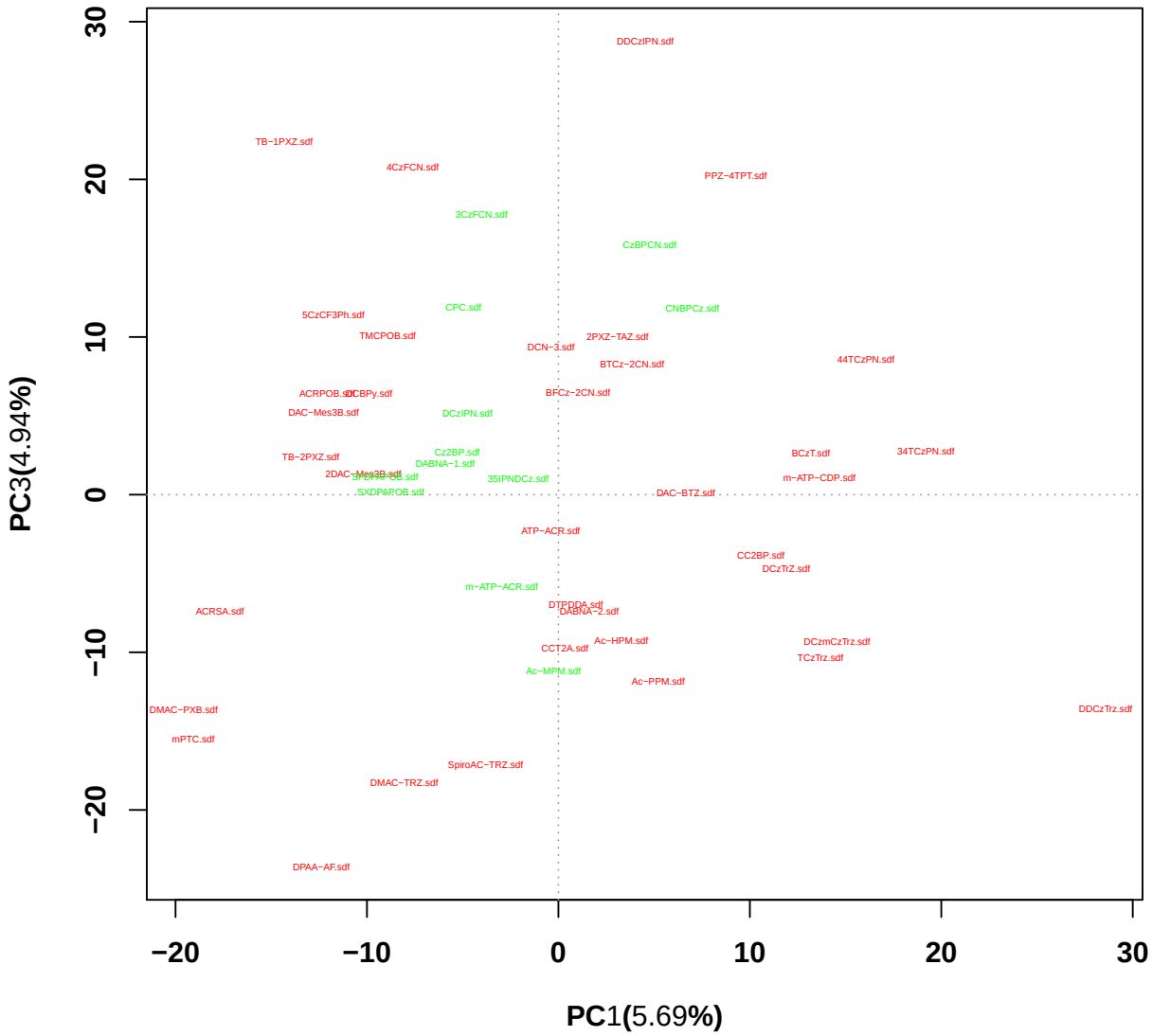


Figure C.8: Scores plot from the PLSR when the eleven potential outliers are removed. All the descriptors are included and no variable selection is performed. PC 3 is plotted against PC 1 and the descriptor variance explained by each of these PCs is given in parenthesis.

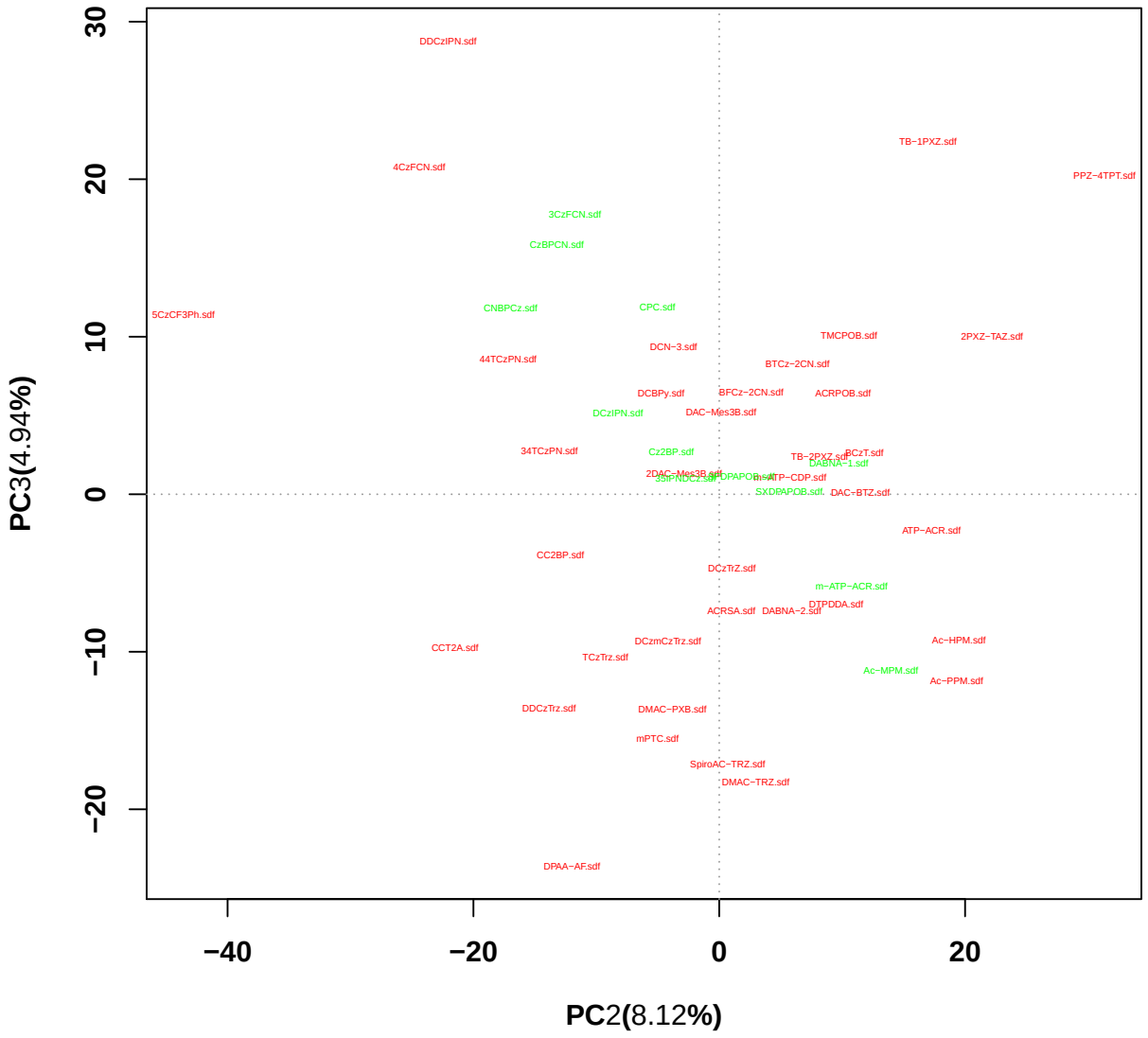


Figure C.9: Scores plot from the PLSR when the eleven potential outliers are removed. All the descriptors are included and no variable selection is performed. PC 3 is plotted against PC 2 and the descriptor variance explained by each of these PCs is given in parenthesis.

C.3 Solvent effects

Table C.2: Observed and predicted ΔE_{st} values for the test set, ΔE_{st}^{obs} and ΔE_{st}^{pred}, respectively. Only the objects with experimental ΔE_{st} values measured in toluene are considered. All descriptors are included and no variable selection is performed. Note that this is the third root of the experimental ΔE_{st} values.

Structure	ΔE_{st}^{obs}	ΔE_{st}^{pred}
1	0.81	0.50
2	0.77	0.68
3CzFCN	0.39	0.46
Cz2BP	0.59	0.44
DCBPy	0.41	0.49
DPXZPO	0.58	0.56
DTPDDA	0.52	0.53
m-ATP-ACR	0.51	0.49

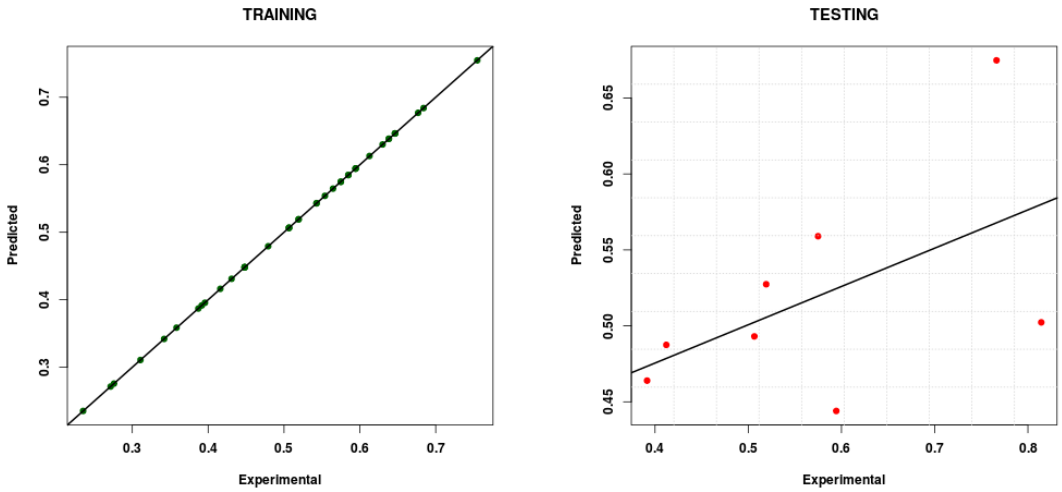


Figure C.10: Fit of the predicted values in the training set and the testing set when all the descriptors are included. Only objects with experimental ΔE_{st} values measured in toluene are considered. The y- and x-axis represents the predicted and the experimental ΔE_{st} values, respectively.

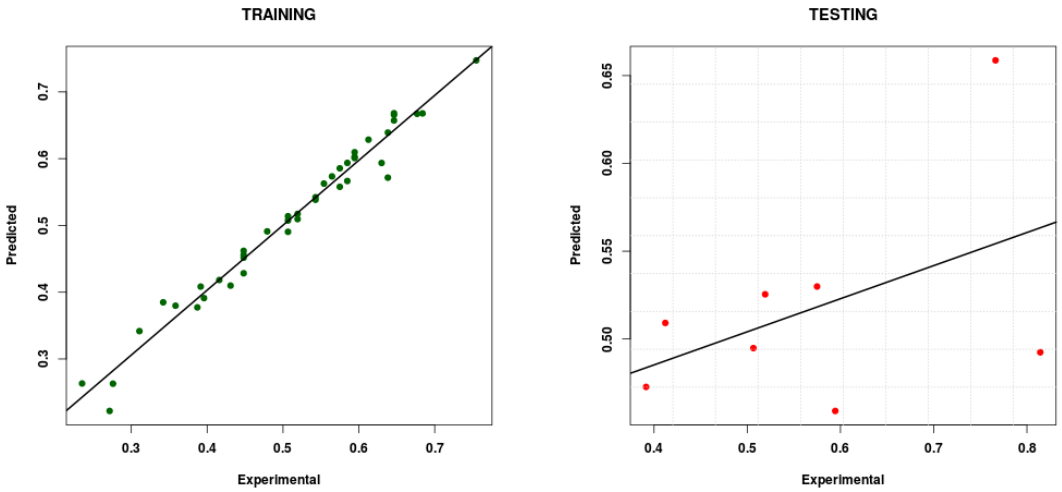


Figure C.11: Fit of the predicted values in the training set and the testing set when all the descriptors are included and variable selection is performed with VIP=0.8. Only objects with experimental ΔE_{st} values measured in toluene are considered. The y- and x-axis represents the predicted and the experimental ΔE_{st} values, respectively.

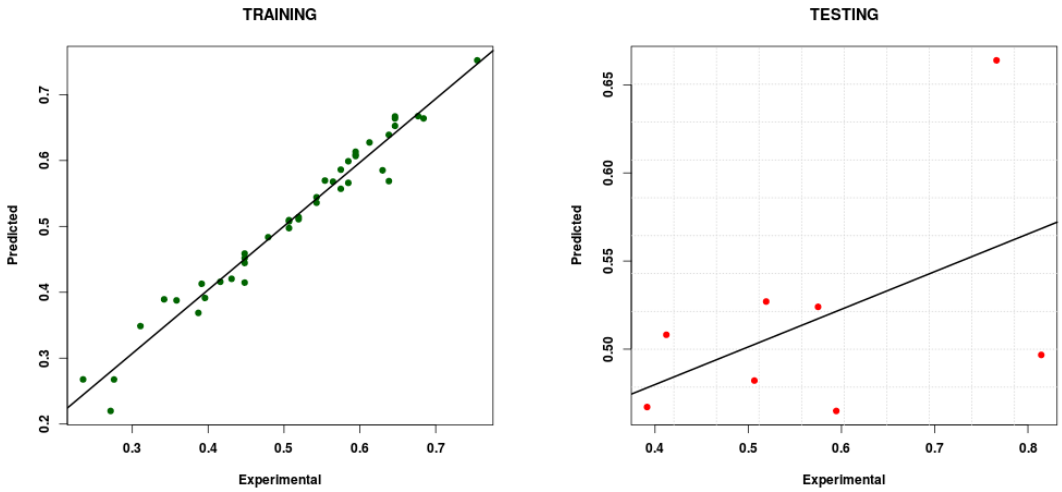


Figure C.12: Fit of the predicted values in the training set and the testing set when all the descriptors are included and variable selection is performed with VIP=1. Only objects with experimental ΔE_{st} values measured in toluene are considered. The y- and x-axis represents the predicted and the experimental ΔE_{st} values, respectively.

D Multiple conformations

The Boltzmann distribution can be regarded as a probability distribution of possible states for a molecule [164]. The probability of finding a molecule in state i with energy ε_i is given by

$$P_i = \frac{\exp -\varepsilon/k_b T}{\sum_{i=1}^{n} \exp \varepsilon_i / k_B T} \tag{D.1}$$

where k_B is the Boltzmann constant, T is the temperature and n is the number of states accessible to the system. The numerator represents the weight of each state and the denominator is the sum of weights, also knows as the partition function.

Table D.1: Results for the training and testing of the PLSR model when multiple conformations are considered. O denotes the result obtain with the original data set, M for when the mean of the multiple conformations are used and B for when the Boltzmann weights are used. All of the descriptors are included. NC is the number of principal components, NV is the number of variables, R^2_{CV} is the correlation coefficient of the cross-validated training set and R^2_{test} is the correlation coefficient of the test set.

				Training			Testing		
	Descriptor	NV	NC	R^2_{CV}	RMSEP	MAE	R^2_{test}	RMSEP	MAE
O	NOVARSEL	3554	3	0.08	0.13	0.02	0.17	0.12	0.09
	VIP=0.8	1823	3	0.62	0.08	0.02	0.17	0.12	0.09
	VIP=1	1174	3	0.70	0.07	0.02	0.17	0.13	0.10
M	NOVARSEL	3558	3	0.14	0.13	0.02	0.18	0.12	0.09
	VIP=0.8	1849	3	0.62	0.08	0.02	0.18	0.12	0.09
	VIP=1	1178	3	0.70	0.07	0.02	0.19	0.12	0.09
B	NOVARSEL	3550	3	0.15	0.12	0.02	0.18	0.12	0.09
	VIP=0.8	1837	3	0.60	0.09	0.02	0.19	0.12	0.09
	VIP=1	1178	3	0.69	0.08	0.02	0.18	0.12	0.09

Table D.2: Observed and predicted ΔE_{st} values for the test set when multiple conformations are considered. ΔE_{st}^{obs} is the third root of the experimental ΔE_{st} values, ΔE_{st}^{O} denotes the predicted values of the original data, ΔE_{st}^{M} denotes the predicted values when the mean of the multiple conformations is used and ΔE_{st}^{B} denotes the predicted values when the Boltzmann weights are used. All descriptors are included.

		NOVARSEL			VIP=0.8			VIP=1		
Molecule	ΔE_{st}^{obs}	ΔE_{st}^{O}	ΔE_{st}^{M}	ΔE_{st}^{B}	ΔE_{st}^{O}	ΔE_{st}^{M}	ΔE_{st}^{B}	ΔE_{st}^{O}	ΔE_{st}^{M}	ΔE_{st}^{B}
2	0.77	0.76	0.75	0.75	0.77	0.77	0.77	0.79	0.80	0.79
35IPNDCz	0.52	0.43	0.44	0.43	0.42	0.42	0.42	0.41	0.42	0.41
3CzFCN	0.39	0.46	0.46	0.46	0.45	0.45	0.45	0.44	0.44	0.44
3	0.68	0.65	0.65	0.65	0.65	0.64	0.63	0.65	0.65	0.63
Ac-MPM	0.58	0.52	0.52	0.52	0.52	0.52	0.52	0.51	0.51	0.51
BTCz-2CN	0.55	0.49	0.49	0.49	0.49	0.50	0.50	0.50	0.50	0.50
CC2BP	0.52	0.58	0.57	0.57	0.59	0.58	0.58	0.59	0.57	0.57
Cz2BP	0.59	0.45	0.44	0.45	0.44	0.44	0.44	0.43	0.43	0.43
DAC-BTZ	0.59	0.50	0.50	0.50	0.51	0.51	0.51	0.51	0.52	0.51
DAC-Mes3B	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.41	0.40
DCzIPN	0.37	0.45	0.45	0.45	0.42	0.45	0.45	0.45	0.45	0.45
DMAC-DPS	0.43	0.55	0.54	0.54	0.56	0.55	0.54	0.57	0.57	0.56
DTC-mBPSB	0.62	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.70	0.70
DTC-pBPSB	0.37	0.70	0.70	0.70	0.70	0.70	0.70	0.71	0.71	0.71
DTPDDA	0.52	0.49	0.51	0.51	0.50	0.50	0.50	0.50	0.51	0.51
SPXZPO	0.64	0.43	0.45	0.45	0.42	0.45	0.45	0.43	0.45	0.45
TB-1PXZ	0.49	0.36	0.36	0.36	0.37	0.37	0.38	0.38	0.38	0.38

E Cubist result

Table E.1: Observed and predicted ΔE_{st} values from the three Cubist calculations. $\Delta E_{st}^{pred,a}$ denotes the prediction with the full data set, $\Delta E_{st}^{pred,b}$ where the eleven objects are exluded and $\Delta E_{st}^{pred,c}$ when only toluene as solvent is considered.

MOLECULE	ΔE_{st}^{obs}	$\Delta E_{st}^{pred,a}$	$\Delta E_{st}^{pred,b}$	$\Delta E_{st}^{pred,c}$
1	0.54	0.15	-	0.46
2DAC-Mes3B	0.06	0.06	0.06	0.07
2PXZ-TAZ	0.23	0.17	0.22	0.22
2	0.45	0.15	-	0.46
34TCzPN	0.21	0.24	0.22	0.22
35IPNDCz	0.14	0.14	0.14	0.14
3CzFCN	0.06	0.07	0.07	0.07
3	0.32	0.12	-	0.31
44TCzPN	0.21	0.24	0.22	0.22
4CzFCN	0.06	0.10	0.08	0.07
5CzCF3Ph	0.02	0.02	0.03	0.03
Ac-HPM	0.18	0.19	0.18	0.18
Ac-MPM	0.19	0.20	0.18	0.18
Ac-PPM	0.19	0.19	0.19	0.18
ACRPOB	0.09	0.08	0.09	0.09
ACRSA	0.03	0.03	0.03	0.03
ATP-ACR	0.16	0.17	0.16	0.16
BCzT	0.31	0.23	0.28	0.29
BFCz-2CN	0.13	0.14	0.13	0.14
BTCz-2CN	0.17	0.13	0.16	0.16
CC2BP	0.14	0.17	0.15	0.16
CCT2A	0.06	0.10	0.09	-
CNBPCz	0.27	0.24	0.25	0.26
CPC	0.04	0.06	0.04	0.04

Cz2BP	0.21	0.16	0.20	0.19
CzAcSF	0.14	0.07	-	-
CzBPCN	0.27	0.24	0.26	0.26
DABNA-1	0.18	0.09	0.15	-
DABNA-2	0.14	0.16	0.14	-
DAC-BTZ	0.20	0.23	0.21	0.20
DAC-Mes3B	0.06	0.06	0.06	0.08
DCBPy	0.07	0.10	0.08	0.08
DCN-3	0.13	0.17	0.14	0.14
DCzIPN	0.05	0.11	0.05	-
DCzmCzTrz	0.20	0.20	0.20	0.20
DCzTrZ	0.25	0.20	0.21	0.23
DDCzIPN	0.13	0.11	0.14	0.14
DDCzTrz	0.27	0.24	0.26	0.26
DMAC-DPS	0.08	0.06	-	0.08
DMAC-PXB	0.01	0.06	0.02	0.02
DMAC-TRZ	0.05	0.11	0.06	0.05
DMOC-DPS	0.21	0.39	-	0.21
DPAA-AF	0.02	0.12	0.05	0.04
DPXZPO	0.19	0.17	-	0.19
DTC-mBPSB	0.24	0.01	-	-
DTC-pBPSB	0.05	0.25	-	-
DTPDDA	0.14	0.14	0.14	0.14
m-ATP-ACR	0.13	0.12	0.13	0.13
m-ATP-CDP	0.26	0.26	0.26	0.26
mPTC	0.01	0.03	0.02	-
PPZ-4TPT	0.43	0.26	0.38	0.42
SFDPAPOB	0.09	0.09	0.09	0.09
SpiroAC-TRZ	0.07	0.12	0.10	0.10
SPXZPO	0.26	0.15	-	0.25
SXDPAPOB	0.09	0.09	0.09	0.09

TB-1PXZ	0.12	0.06	0.10	-
TB-2PXZ	0.05	0.06	0.05	-
TCzTrz	0.16	0.17	0.17	0.18
TMCPOB	0.09	0.07	0.09	0.09
TPXZPO	0.11	0.13	-	0.11

F DFT output coordinates

Table F.1: Output coordinates for structure 1 from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
C	-2.186005	-6.863399	-0.40596
C	-1.023608	-9.214837	-0.008476
C	-2.171668	-11.011715	1.535263
C	-4.505367	-10.513467	2.667223
C	-5.678217	-8.186558	2.243171
C	-4.531708	-6.362833	0.727435
N	-0.984923	-5.01897	-1.925831
C	1.564083	-4.407465	-1.556112
C	3.01744	-3.509425	-3.600095
C	5.383784	-2.457621	-3.193807
C	6.363349	-2.370611	-0.740765
C	5.064188	-3.515907	1.25632
C	2.685775	-4.543692	0.854939
S	8.676682	0.00861	0.001313
O	10.059231	0.71859	-2.341184
O	10.06075	-0.701413	2.342903
C	6.363501	2.38741	0.744835
C	5.384079	2.472286	3.198044
C	3.016	3.520031	3.604674
C	1.561311	4.416021	1.560808
C	2.68311	4.556334	-0.849849
C	5.063063	3.532265	-1.251675
N	-0.989622	5.020464	1.929752
C	-2.48504	3.455027	3.516587
C	-2.231535	0.813652	3.441299
C	-3.7306	-0.69823	4.98961
C	-5.516613	0.390202	6.602427
C	-5.778315	3.017081	6.665237
C	-4.266034	4.548274	5.146141
C	-2.193132	6.864032	0.41075
C	-4.533601	6.358157	-0.730979
C	-5.681475	8.181613	-2.246066
C	-4.515058	10.513428	-2.661019
C	-2.18643	11.016786	-1.520909
C	-1.037168	9.220102	0.022122
C	-2.482501	-3.4632	-3.520302
C	-4.252401	-4.568205	-5.153872
C	-5.767142	-3.047243	-6.68077
C	-5.51892	-0.418985	-6.621344
C	-3.744122	0.681263	-5.004177
C	-2.242402	-0.820511	-3.448477
H	0.791419	-9.613997	-0.907062
H	-1.241869	-12.832024	1.83054
H	-5.406009	-11.932655	3.866013
H	-7.500165	-7.767606	3.121448
H	-5.445788	-4.538015	0.424692
H	2.213992	-3.524706	-5.500131
H	6.437482	-1.622985	-4.75957
H	5.847069	-3.466394	3.165078
H	1.635545	-5.331551	2.445112
H	6.438954	1.638304	4.763362
H	2.212009	3.532799	5.504507
H	1.631624	5.343344	-2.439621
H	5.846097	3.484548	-3.160423
H	-0.85745	-0.052637	2.168396
H	-3.520023	-2.751501	4.900433
H	-6.69827	-0.803529	7.803159
H	-7.159391	3.891509	7.927773
H	-4.459083	6.602669	5.202738
H	-5.442383	4.529558	-0.435137
H	-7.499304	7.758705	-3.130962
H	-5.41678	11.932408	-3.859246
H	-1.261764	12.840818	-1.809249
H	0.773902	9.622904	0.927062
H	-4.434841	-6.623656	-5.207295
H	-7.1395	-3.930588	-7.946598
H	-6.70251	0.766748	-7.828088
H	-3.54427	2.735756	-4.917675
H	-0.877029	0.054631	-2.18225

Table F.2: Output coordinates for structure 2DAC-Mes3B from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
N	0.181098	0.686262	0.322232
C	2.275631	2.258923	0.122936
C	1.450345	4.811981	0.125083
C	-1.265248	4.759078	0.306212
C	-1.980767	2.176219	0.430346
C	0.217109	-1.989065	0.303377
C	1.340456	-5.924374	-1.769888
C	1.310876	-3.283594	-1.716555
C	-0.867033	-3.339429	2.291449
C	-0.939029	-5.979794	2.245391
C	0.170922	-7.318479	0.202915
C	2.554107	-7.193571	-4.006232
C	-2.185159	-7.314626	4.424978
B	0.035686	-10.294134	0.083175
C	-2.591112	-11.654831	0.405721
C	2.503131	-11.883196	-0.39579
C	-7.316494	-14.207229	0.911685
C	-7.039404	-12.149022	-0.718565
C	-4.749091	-10.847421	-0.973794
C	-2.856686	-13.755007	2.063249
C	-5.197387	-14.95718	2.304894
C	4.732503	-11.446711	1.040444
C	6.88845	-12.90864	0.601677
C	6.968246	-14.77365	-1.272144
C	4.784603	-15.160969	-2.704378
C	2.56137	-13.786568	-2.28587
C	-4.679027	-8.64835	-2.779467
C	-9.799535	-15.537376	1.210013
C	-0.69381	-14.735013	3.631176
C	0.311305	-14.373962	-3.925986
C	9.318472	-16.296249	-1.700891
C	4.854761	-9.469599	3.082528
C	4.845822	1.655164	-0.005466
C	6.566872	3.631596	-0.176994
C	5.782178	6.187009	-0.153959
C	3.212209	6.77651	0.02286
C	-3.112977	6.646283	0.273089
C	-5.658413	5.953261	0.422196
C	-6.332869	3.371273	0.574666
C	-4.524505	1.467091	0.548097
N	7.607603	8.13714	-0.328365
C	9.830545	8.010653	1.13458
C	7.155983	10.229716	-1.918666
C	6.20656	14.407503	-5.068308
C	7.26918	14.751533	-2.675367
C	7.727361	12.691504	-1.101253
C	6.103476	9.885254	-4.333307
C	5.626633	11.961308	-5.879804
C	9.758948	7.103483	3.631169
C	11.972708	6.942136	5.048073
C	14.286599	7.698121	4.024096
C	14.358706	8.618294	1.548955
C	12.162926	8.762079	0.104952
N	-7.57676	7.820261	0.408764
C	-9.705846	7.512257	-1.164094
C	-7.271112	10.044871	1.846151
C	-9.455325	6.472569	-3.59634
C	-11.575807	6.139246	-5.121496
C	-13.972452	6.851558	-4.271833
C	-14.222651	7.90255	-1.860994
C	-12.120913	8.2185	-0.308714
C	-5.966267	12.132941	5.715355
C	-6.618497	14.483833	4.701882
C	-7.601753	14.601412	2.254002
C	-7.91319	12.412087	0.825372
C	-6.294787	9.92873	4.31452
H	2.12905	-2.209705	-3.278481
H	-1.669902	-2.309369	3.890562
H	3.032719	-5.809239	-5.477667
H	4.30191	-8.175518	-3.454757
H	1.31283	-8.62683	-4.857473
H	-1.075537	-8.940348	5.08836
H	-4.052186	-8.051238	3.877175
H	-2.456796	-6.025776	6.030318
H	-8.666653	-11.528575	-1.835158
H	-5.373093	-16.549752	3.614402
H	8.573618	-12.572418	1.754706
H	4.807324	-16.592418	-4.19797
H	-4.475372	-6.842	-1.769063
H	-3.086045	-8.765618	-4.10867
H	-6.427196	-8.557861	-3.896336
H	-11.11955	-15.048977	-0.317031
H	-10.703446	-15.016856	3.013827
H	-9.555484	-17.602462	1.230995
H	-1.372184	-16.069133	5.071034
H	0.325093	-13.210192	4.60801
H	0.706463	-15.725865	2.455766
H	0.860129	-15.579093	-5.52605
H	-0.572705	-12.658923	-4.697091
H	-1.166069	-15.372746	-2.855872
H	9.117273	-17.584545	-3.317055
H	10.957588	-15.063728	-2.055423
H	9.778792	-17.441437	-0.023817
H	6.58712	-9.674888	4.209074
H	4.842242	-7.551288	2.279545
H	3.237324	-9.576851	4.382546
H	5.489268	-0.304172	-0.00311
H	8.581685	3.211749	-0.318862
H	2.59053	8.743643	0.061195
H	-2.578977	8.634054	0.127622
H	-8.330756	2.870796	0.692578
H	-5.079974	-0.517226	0.637421
H	5.835354	16.029679	-6.290004
H	7.71745	16.653574	-2.004951
H	8.528407	12.973059	0.779873
H	5.654515	7.977076	-4.979073
H	4.811593	11.658442	-7.753747
H	7.955138	6.521972	4.448142
H	11.880116	6.238405	6.988085
H	16.016094	7.576966	5.14469
H	16.155085	9.203055	0.712927
H	12.236138	9.453836	-1.838025
H	-7.585371	5.926502	-4.277782
H	-11.343897	5.334264	-7.009903
H	-15.628935	6.59547	-5.476927
H	-16.086231	8.454949	-1.160825
H	-12.331076	9.011993	1.584979
H	-5.211712	12.007124	7.634218
H	-6.364684	16.207391	5.80936
H	-8.10339	16.426117	1.425383
H	-8.653011	12.519906	-1.098072
H	-5.789304	8.096478	5.109033

Table F.3: Output coordinates for structure 2PXZ-TAZ from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
O	1.443419	17.459207	0.107448
C	2.036292	15.877167	2.094266
C	3.58247	16.857051	3.973386
C	4.274522	15.361495	6.041844
C	3.402458	12.88472	6.193855
C	1.84761	11.895453	4.299529
C	1.134146	13.373991	2.223678
N	-0.423427	12.469447	0.28687
C	-1.104293	9.864174	0.220328
C	0.428346	8.148615	-1.074158
C	-0.206928	5.599804	-1.138239
C	-2.379367	4.725667	0.116321
C	-3.922286	6.474227	1.388835
C	-3.29127	9.026194	1.437785
C	-3.149454	2.071047	0.086881
N	-5.530531	1.286485	0.100915
N	-5.535835	-1.272774	0.045086
C	-3.158006	-2.065777	0.003201
N	-1.571058	-0.000322	0.026638
C	1.124563	-0.005543	-0.005317
C	2.394036	-0.868923	-2.151134
C	5.028198	-0.877052	-2.168007
C	6.374126	-0.016186	-0.067013
C	5.081487	0.850259	2.064922
C	2.447669	0.853024	2.10976
C	-2.397365	-4.722749	-0.047104
C	-0.204487	-5.60575	1.164768
C	0.42247	-8.155997	1.082263
C	-1.138115	-9.863851	-0.189071
C	-3.345582	-9.017128	-1.36262
C	-3.968888	-6.463553	-1.294688
N	-0.461526	-12.469618	-0.280733
C	1.063556	-13.362829	-2.248626
C	1.73261	-11.875742	-4.333061
C	3.255617	-12.853357	-6.259082
C	4.140134	-15.327013	-6.130368
C	3.493325	-16.830734	-4.053261
C	1.978747	-15.862264	-2.14269
O	1.431129	-17.45144	-0.148761
C	-0.034182	-16.523346	1.799201
C	-1.002672	-14.042789	1.776609
C	-2.488026	-13.262001	3.82237
C	-2.985362	-14.901539	5.833901
C	-2.006583	-17.342974	5.835213
C	-0.52435	-18.149072	3.798433
C	-1.013067	14.035862	-1.762259
C	-0.056485	16.520786	-1.809022
C	-0.593822	18.140215	-3.801231
C	-2.112654	17.323213	-5.806446
C	-3.080278	14.877493	-5.78077
C	-2.535469	13.244198	-3.776393
H	4.233153	18.806163	3.782483
H	5.488258	16.148813	7.512807
H	3.922087	11.687258	7.7924
H	1.173747	9.951726	4.424962
H	2.115685	8.831836	-2.045477
H	0.991953	4.296123	-2.189817
H	-5.622709	5.804295	2.34441
H	-4.482179	10.386374	2.431977
H	1.315732	-1.52663	-3.782345
H	6.033004	-1.554007	-3.839371
H	8.438589	-0.020764	-0.091338
H	6.127502	1.522806	3.712621
H	1.410304	1.515074	3.765516
H	1.017372	-4.307452	2.196523
H	2.125984	-8.846362	2.019708
H	-4.557975	-10.371324	-2.338848
H	-5.684965	-5.786414	-2.216616
H	1.048918	-9.934388	-4.440511
H	3.740321	-11.649396	-7.863684
H	5.329145	-16.105727	-7.625898
H	4.155899	-18.777464	-3.879782
H	-3.257848	-11.34993	3.831296
H	-4.151088	-14.237815	7.402263
H	-2.382994	-18.632881	7.400688
H	0.276415	-20.049847	3.728797
H	0.199824	20.044573	-3.75111
H	-2.526178	18.608347	-7.366465
H	-4.274335	14.205562	-7.32412
H	-3.29691	11.328812	-3.776339

Table F.4: Output coordinates for structure 3 from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
C	8.675116	2.224733	-0.892735
C	6.788521	3.745155	-1.941562
C	7.289449	4.903834	-4.247546
C	9.597069	4.599432	-5.568278
C	11.43543	3.036736	-4.495618
C	10.991967	1.846388	-2.172463
N	8.690391	0.809113	1.333123
C	6.67435	0.620688	3.061934
C	5.730282	-1.762768	3.73783
C	3.722324	-1.954021	5.42021
C	2.644975	0.247008	6.400133
C	3.568572	2.630783	5.750926
C	5.601255	2.812758	4.093949
S	0	0	8.525109
O	0.225387	-2.43876	9.902443
O	-0.225387	2.43876	9.902443
C	-2.644975	-0.247008	6.400133
C	-3.722324	1.954021	5.42021
C	-5.730282	1.762768	3.73783
C	-6.67435	-0.620688	3.061934
C	-5.601255	-2.812758	4.093949
C	-3.568572	-2.630783	5.750926
N	-8.690391	-0.809113	1.333123
C	-10.991835	0.468968	1.493972
C	-12.467199	-0.134667	-0.651938
C	-14.896396	0.883404	-0.921911
C	-15.884965	2.488342	0.926746
C	-14.367285	3.013736	3.06833
C	-11.948867	2.035753	3.391993
C	-8.675116	-2.224733	-0.892735
C	-6.788521	-3.745155	-1.941562
C	-7.289449	-4.903834	-4.247546
C	-9.597069	-4.599432	-5.568278
C	-11.43543	-3.036736	-4.495618
C	-10.991967	-1.846388	-2.172463
C	10.991835	-0.468968	1.493972
C	12.467199	0.134667	-0.651938
C	14.896396	-0.883404	-0.921911
C	15.884965	-2.488342	0.926746
C	14.367285	-3.013736	3.06833
C	11.948867	-2.035753	3.391993
C	18.526096	-3.671301	0.727059
C	20.148203	-2.781937	2.969125
C	18.285695	-6.567481	0.784373
C	19.876499	-2.923674	-1.722161
C	9.982252	5.975045	-8.092564
C	12.563356	5.399194	-9.270432
C	9.801495	8.842226	-7.650714
C	7.921944	5.145448	-9.967457
C	-9.982252	-5.975045	-8.092564
C	-18.526096	3.671301	0.727059
C	-20.148203	2.781937	2.969125
C	-18.285695	6.567481	0.784373
C	-19.876499	2.923674	-1.722161
C	-7.921944	-5.145448	-9.967457
C	-12.563356	-5.399194	-9.270432
C	-9.801495	-8.842226	-7.650714
H	4.973805	4.016634	-1.001
H	5.813271	6.095469	-5.059547
H	13.237353	2.708092	-5.439892
H	6.565062	-3.460331	2.916461
H	2.985124	-3.805993	5.951768
H	2.734958	4.325169	6.580735
H	6.38187	4.658891	3.606877
H	-2.985124	3.805993	5.951768
H	-6.565062	3.460331	2.916461
H	-6.38187	-4.658891	3.606877
H	-2.734958	-4.325169	6.580735
H	-15.986633	0.388618	-2.599007
H	-15.110423	4.235434	4.556296
H	-10.855285	2.47123	5.085128
H	-4.973805	-4.016634	-1.001
H	-5.813271	-6.095469	-5.059547
H	-13.237353	-2.708092	-5.439892
H	15.986633	-0.388618	-2.599007
H	15.110423	-4.235434	4.556296
H	10.855285	-2.47123	5.085128
H	22.057453	-3.604261	2.853369
H	19.328473	-3.348016	4.793796
H	20.336794	-0.709451	2.970297
H	20.164262	-7.451542	0.628582
H	17.118623	-7.246417	-0.799376
H	17.418775	-7.241776	2.549417
H	21.748405	-3.827183	-1.7904
H	18.82039	-3.536672	-3.405843
H	20.166684	-0.868214	-1.851336
H	12.753708	6.428693	-11.067368
H	14.129248	5.995206	-8.038197
H	12.794429	3.372706	-9.678384
H	10.080431	9.869418	-9.440234
H	11.258235	9.478526	-6.307525
H	7.947129	9.393127	-6.889177
H	8.180588	6.104786	-11.797151
H	6.014755	5.60644	-9.280758
H	8.00267	3.09229	-10.297321
H	-22.057453	3.604261	2.853369
H	-19.328473	3.348016	4.793796
H	-20.336794	0.709451	2.970297
H	-20.164262	7.451542	0.628582
H	-17.118623	7.246417	-0.799376
H	-17.418775	7.241776	2.549417
H	-21.748405	3.827183	-1.7904
H	-18.82039	3.536672	-3.405843
H	-20.166684	0.868214	-1.851336
H	-8.180588	-6.104786	-11.797151
H	-8.00267	-3.09229	-10.297321
H	-6.014755	-5.60644	-9.280758
H	-12.753708	-6.428693	-11.067368
H	-12.794429	-3.372706	-9.678384
H	-14.129248	-5.995206	-8.038197
H	-10.080431	-9.869418	-9.440234
H	-11.258235	-9.478526	-6.307525
H	-7.947129	-9.393127	-6.899177

Table F.5: Output coordinates for structure BCzT from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
C	-0.026657	6.668912	-1.869378
C	1.10847	4.30547	-1.959481
C	3.168944	3.761637	-0.384947
C	4.097695	5.626688	1.251441
C	2.978641	8.00042	1.304344
C	0.890102	8.553089	-0.238732
N	4.280942	1.337645	-0.443023
C	-0.329528	11.060471	-0.138043
N	0.821404	12.88233	1.203181
N	-2.519839	11.356002	-1.386196
C	-3.563846	13.66355	-1.226105
N	-2.577754	15.597635	0.08863
C	-0.385164	15.113994	1.271457
C	0.778711	17.176609	2.755893
C	-5.964872	14.10303	-2.587354
C	0.629133	21.44064	4.437384
C	-0.459204	19.51114	3.016687
C	3.131225	16.82131	3.935456
C	4.214447	18.758803	5.348886
C	2.967998	21.07238	5.606826
C	-7.114168	12.138272	-3.955283
C	-7.122996	16.492	-2.534932
C	-9.378246	16.903582	-3.826515
C	-10.506593	14.942158	-5.188136
C	-9.368995	12.559687	-5.244862
C	2.983021	-0.935552	-0.155822
C	4.713994	-2.962494	-0.421263
C	7.160468	-1.854553	-0.873988
C	6.825013	0.808018	-0.876561
C	-0.3505	-3.870775	0.635717
C	1.317475	-5.940007	0.329527
C	3.869615	-5.453893	-0.181659
C	0.441295	-1.374913	0.408933
C	9.564168	-2.871751	-1.305877
C	11.581071	-1.237407	-1.737487
C	11.209786	1.394084	-1.763708
C	8.835505	2.45569	-1.345845
C	0.369217	-8.56376	0.525776
C	0.200077	-12.86854	-1.024641
C	-1.498868	-13.499921	0.951541
C	-2.295242	-11.690026	2.702013
C	-1.344014	-9.251778	2.463177
C	1.131297	-10.404571	-1.218038
C	-0.963841	-17.034471	-1.386298
N	-2.179567	-16.025344	0.714764
C	0.53812	-15.118878	-2.518447
C	0.390833	-19.997419	-4.50586
C	-1.038782	-19.492017	-2.352582
C	1.962176	-15.680701	-4.677031
C	1.875087	-18.119224	-5.662753
C	-3.83158	-17.355494	2.350088
C	-6.030434	-18.465001	1.385673
C	-7.640262	-19.789164	2.997755
C	-7.084997	-19.980284	5.571165
C	-4.897998	-18.854263	6.527853
C	-3.258724	-17.555867	4.924938
H	-1.6244	7.092135	-3.100944
H	0.418479	2.86668	-3.266262
H	5.679993	5.192343	2.501407
H	3.687983	9.446028	2.591408
H	-0.354424	23.245455	4.633092
H	-2.285334	19.784772	2.09983
H	4.09379	15.010026	3.727878
H	6.045689	18.463388	6.255943
H	3.82054	22.589727	6.719138
H	-6.217346	10.282031	-3.986207
H	-6.228614	18.014936	-1.471561
H	-10.26231	18.768758	-3.774527
H	-12.275438	15.269356	-6.203192
H	-10.24744	11.018454	-6.301641
H	-2.344419	-4.23998	1.020617
H	5.209173	-7.015803	-0.354873
H	-0.891515	0.180366	0.646032
H	9.838504	-4.918575	-1.308276
H	13.469238	-2.002532	-2.068698
H	12.813933	2.642674	-2.1252
H	8.55785	4.498303	-1.388634
H	-3.611312	-12.173389	4.214673
H	-1.921305	-7.82575	3.839363
H	2.407516	-9.908419	-2.763761
H	0.360507	-21.901475	-5.303959
H	-2.179228	-20.957797	-1.455816
H	3.122472	-14.22069	-5.564095
H	2.971856	-18.584845	-7.348278
H	-6.463967	-18.275442	-0.622972
H	-9.352153	-20.657199	2.236472
H	-8.357568	-21.006776	6.832339
H	-4.44785	-19.004344	8.537581
H	-1.528393	-16.697566	5.641098

Table F.6: Output coordinates for structure BTCz-2CN from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
C	-1.072388	-0.802447	-9.508
C	1.072388	0.802447	-9.508
C	2.085363	1.615061	-7.208611
C	1.043716	0.846785	-4.910865
C	-1.043716	-0.846785	-4.910865
C	-2.085363	-1.615061	-7.208611
C	2.189399	1.633326	-11.809947
C	-2.189399	-1.633326	-11.809947
N	2.077955	1.767097	-2.645237
N	-2.077955	-1.767097	-2.645237
C	4.490413	1.165189	-1.736672
C	4.58393	1.785422	0.863
C	2.166276	2.822911	1.549026
C	0.670297	2.81775	-0.670578
C	-0.670297	-2.81775	-0.670578
C	-2.166276	-2.822911	1.549026
C	-4.58393	-1.785422	0.863
C	-4.490413	-1.165189	-1.736672
C	1.125743	3.738896	3.804577
C	-1.348546	4.637472	3.788833
C	-2.763749	4.704568	1.544419
C	-1.76237	3.824452	-0.724986
C	1.76237	-3.824452	-0.724986
C	2.763749	-4.704568	1.544419
C	1.348546	-4.637472	3.788833
C	-1.125743	-3.738896	3.804577
C	-6.500196	0.001516	-3.006126
C	-6.793626	-1.187485	2.19162
C	-8.846625	-0.008548	0.953645
C	-8.666085	0.568435	-1.639966
C	6.500196	-0.001516	-3.006126
C	8.666085	-0.568435	-1.639966
C	8.846625	0.008548	0.953645
C	6.793626	1.187485	2.19162
S	-7.365009	-1.698512	5.40635
C	-10.383876	-0.341313	5.153521
C	-10.921636	0.475614	2.66428
C	10.921636	-0.475614	2.66428
C	10.383876	0.341313	5.153521
S	7.365009	1.698512	5.40635
C	13.268892	-1.613926	2.17954
C	15.000741	-1.914408	4.133061
C	14.429977	-1.089626	6.588981
C	12.116551	0.046544	7.122449
C	-13.268892	1.613926	2.17954
C	-12.116551	-0.046544	7.122449
C	-14.429977	1.089626	6.588981
C	-15.000741	1.914408	4.133061
N	-3.132824	-2.322801	-13.686776
N	3.132824	2.322801	-13.686776
H	3.692493	2.905659	-7.198911
H	-3.692493	-2.905659	-7.198911
H	2.22346	3.744112	5.550979
H	-2.194681	5.322118	5.541953
H	-4.690604	5.443902	1.568812
H	-2.854005	3.885768	-2.471144
H	2.854005	-3.885768	-2.471144
H	4.690604	-5.443902	1.568812
H	2.194681	-5.322118	5.541953
H	-2.22346	-3.744112	5.550979
H	-6.364954	0.498852	-5.0009
H	-10.252919	1.484653	-2.588322
H	6.364954	-0.498852	-5.0009
H	10.252919	-1.484653	-2.588322
H	13.719896	-2.255957	0.269589
H	16.826788	-2.798788	3.752324
H	15.810752	-1.335855	8.103888
H	11.669277	0.690224	9.031488
H	-13.719896	2.255957	0.269589
H	-11.669277	-0.690224	9.031488
H	-15.810752	1.335855	8.103888
H	-16.826788	2.798788	3.742324

Table F.7: Output coordinates for structure Cz2BP from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
C	4.48575	10.571718	-0.559983
C	6.174775	12.417613	-1.381715
C	5.329571	14.714139	-2.418274
C	2.756645	15.205362	-2.668411
C	1.019122	13.382219	-1.857204
C	-1.699395	13.251955	-1.873379
C	-3.596917	14.929754	-2.640281
C	-6.115114	14.227164	-2.343243
C	-6.748806	11.881037	-1.267665
C	-4.898334	10.177202	-0.487314
C	-2.373081	10.878382	-0.821639
N	-0.170435	9.585933	-0.175441
C	1.904514	11.086758	-0.789212
C	-0.072744	7.187209	0.989945
C	1.205922	6.881118	3.291594
C	-1.285166	5.119082	-0.13059
C	-1.213721	2.76536	1.038988
C	0.05086	2.439714	3.347995
C	1.24642	4.532124	4.458706
C	0	0	4.762811
C	-0.05086	-2.439714	3.347995
O	0	0	7.090973
C	-1.205922	-6.881118	3.291594
C	-1.24642	-4.532124	4.458706
C	1.213721	-2.76536	1.038988
C	1.285166	-5.119082	-0.13059
C	0.072744	-7.187209	0.989945
N	0.170435	-9.585933	-0.175441
C	-1.904514	-11.086758	-0.789212
C	-1.019122	-13.382219	-1.857204
C	1.699395	-13.251955	-1.873379
C	2.373081	-10.878382	-0.821639
C	-4.48575	-10.571718	-0.559983
C	-6.174775	-12.417613	-1.381715
C	-5.329571	-14.714139	-2.418274
C	-2.756645	-15.205362	-2.668411
C	6.115114	-14.227164	-2.343243
C	6.748806	-11.881037	-1.267665
C	4.898334	-10.177202	-0.487314
C	3.596917	-14.929754	-2.640281
H	5.158937	8.789464	0.229959
H	8.202033	12.061634	-1.21756
H	6.708931	16.118834	-3.038366
H	2.088664	16.977141	-3.492346
H	-3.094916	16.761927	-3.449677
H	-7.616747	15.513262	-2.936069
H	-8.736724	11.376226	-1.028936
H	-5.404211	8.36748	0.361984
H	2.125328	8.504475	4.172078
H	-2.273021	5.364499	-1.924794
H	-2.183215	1.173518	0.15396
H	2.191922	4.295579	6.277737
H	-2.125328	-8.504475	4.172078
H	-2.191922	-4.295579	6.277737
H	2.183215	-1.173518	0.15396
H	2.273021	-5.364499	-1.924794
H	-5.158937	-8.789464	0.229959
H	-8.202033	-12.061634	-1.21756
H	-6.708931	-16.118834	-3.038366
H	-2.088664	-16.977141	-3.492346
H	7.616747	-15.513262	-2.936069
H	8.736724	-11.376226	-1.028936
H	5.404211	-8.36748	0.361984
H	3.094916	-16.761927	-3.459677

Table F.8: Output coordinates for structure SPXZPO from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
C	2.716328	7.074329	0.459772
C	1.993844	8.344205	2.668411
C	3.832036	9.246156	4.333369
C	6.391851	8.897558	3.79978
C	7.118765	7.649539	1.588274
C	5.292448	6.743577	-0.080501
P	0.447624	5.799874	-1.81394
O	1.070717	6.436869	-4.547362
C	-2.666324	6.940624	-0.846483
C	-4.034027	5.857499	1.148094
C	-6.384669	6.844767	1.82116
C	-7.382629	8.906169	0.506811
C	-6.031516	9.975131	-1.492411
C	-3.678607	8.99776	-2.17264
C	0.414149	2.39129	-1.207779
C	-0.286379	0.794976	-3.201501
C	-0.381636	-1.809929	-2.838363
C	0.216195	-2.825552	-0.474787
C	0.928141	-1.248932	1.519478
C	1.036919	1.357346	1.149465
N	0.114018	-5.4924	-0.107841
C	2.26218	-6.969432	-0.549174
C	4.603882	-5.93278	-1.217007
C	6.700555	-7.487986	-1.630847
C	6.490084	-10.099544	-1.395019
C	4.158296	-11.159925	-0.734665
C	2.07836	-9.616596	-0.317357
O	-0.158029	-10.794615	0.324396
C	-2.248675	-9.305057	0.77988
C	-4.465931	-10.539933	1.444035
C	-6.667645	-9.154727	1.924785
C	-6.609158	-6.531191	1.740753
C	-4.375078	-5.285495	1.073802
C	-2.162087	-6.649828	0.576359
H	-0.005426	8.633938	3.093717
H	3.255477	10.236346	6.052192
H	7.826509	9.607569	5.105472
H	9.121379	7.383941	1.156876
H	5.867259	5.77765	-1.813803
H	-3.270164	4.236971	2.174679
H	-7.445581	5.993236	3.375639
H	-9.226477	9.673605	1.034661
H	-6.815226	11.576441	-2.535449
H	-2.625057	9.815775	-3.749628
H	-0.737118	1.600833	-5.048992
H	-0.916646	-3.075393	-4.378546
H	1.404627	-2.081987	3.346779
H	1.618419	2.585799	2.704009
H	4.779005	-3.888331	-1.415494
H	8.502536	-6.623372	-2.145678
H	8.118542	-11.325114	-1.715769
H	3.913781	-13.199409	-0.532554
H	-4.433028	-12.599868	1.569557
H	-8.406096	-10.140369	2.438014
H	-8.307865	-5.417932	2.107863
H	-4.343129	-3.229512	0.916097

Table F.9: Output coordinates for structure TCzTrz from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
C	0.856333	4.488878	11.034203
C	1.185178	6.774004	9.720314
C	1.613381	9.012028	11.036322
C	1.724161	9.001253	13.674928
C	1.407627	6.735147	14.993841
C	0.974819	4.492243	13.687607
C	0.394319	2.112318	9.641807
N	0.397043	2.21114	7.103649
C	0	0	5.933267
N	-0.397043	-2.21114	7.103649
C	-0.394319	-2.112318	9.641807
N	0	0	10.991363
C	-0.856333	-4.488878	11.034203
C	-0.974819	-4.492243	13.687607
C	-1.407627	-6.735147	14.993841
C	-1.724161	-9.001253	13.674928
C	-1.613381	-9.012028	11.036322
C	-1.185178	-6.774004	9.720314
C	0	0	3.138428
C	0.358143	2.25193	1.800519
C	0.35298	2.272283	-0.831062
C	0	0	-2.199131
C	-0.35298	-2.272283	-0.831062
C	-0.358143	-2.25193	1.800519
N	-0.711404	-4.585549	-2.107826
C	0.919823	-6.650316	-1.959826
C	3.077362	-6.981918	-0.476031
C	4.425236	-9.225711	-0.771693
C	3.651028	-11.092479	-2.496969
C	1.499253	-10.745997	-3.972815
C	0.112528	-8.507742	-3.715488
C	-2.117272	-7.530299	-4.936779
C	-3.762708	-8.491936	-6.772868
C	-5.833755	-7.045657	-7.510045
C	-6.297549	-4.675855	-6.401641
C	-4.695627	-3.682714	-4.564734
C	-2.585827	-5.111063	-3.883977
N	0	0	-4.852951
C	-1.693363	1.317202	-6.391306
C	-3.778969	2.780466	-5.716216
C	-5.225576	3.813133	-7.658282
C	-4.622521	3.387458	-10.209629
C	-2.561221	1.889157	-10.869104
C	-1.078325	0.832058	-8.950682
C	1.078325	-0.832058	-8.950682
C	2.561221	-1.889157	-10.869104
C	4.622521	-3.387458	-10.209629
C	5.225576	-3.813133	-7.658282
C	3.778969	-2.780466	-5.716216
C	1.693363	-1.317202	-6.391306
N	0.711404	4.585549	-2.107826
C	-0.919823	6.650316	-1.959826
C	-3.077362	6.981918	-0.476031
C	-4.425236	9.225711	-0.771693
C	-3.651028	11.092479	-2.496969
C	-1.499253	10.745997	-3.972815
C	-0.112528	8.507742	-3.715488
C	2.117272	7.530299	-4.936779
C	3.762708	8.491936	-6.772868
C	5.833755	7.045657	-7.510045
C	6.297549	4.675855	-6.401641
C	4.695627	3.682714	-4.564734
C	2.585827	5.111063	-3.883977
H	1.098313	6.771243	7.66058
H	1.859748	10.778719	9.996887
H	2.058198	10.760724	14.703569
H	1.499051	6.71668	17.056487
H	0.73071	2.716835	14.706512
H	-0.73071	-2.716835	14.706512
H	-1.499051	-6.71668	17.056487
H	-2.058198	-10.760724	14.703569
H	-1.859748	-10.778719	9.996887
H	-1.098313	-6.771243	7.66058
H	0.676532	4.015651	2.815362
H	-0.676532	-4.015651	2.815362
H	3.694029	-5.535864	0.85898
H	6.118275	-9.533901	0.369246
H	4.752513	-12.828598	-2.678366
H	0.89908	-12.184978	-5.32679
H	-3.420485	-10.351006	-7.603509
H	-7.127614	-7.764752	-8.948617
H	-7.94777	-3.581009	-6.984861
H	-5.082977	-1.859828	-3.688812
H	-4.285052	3.088679	-3.744081
H	-6.858276	4.980827	-7.175249
H	-5.793741	4.229408	-11.68612
H	-2.109055	1.522296	-12.850125
H	2.109055	-1.522296	-12.850125
H	5.793741	-4.229408	-11.68612
H	6.858276	-4.980827	-7.175249
H	4.285052	-3.088679	-3.744081
H	-3.694029	5.535864	0.85898
H	-6.118275	9.533901	0.369246
H	-4.752513	12.828598	-2.678366
H	-0.89908	12.184978	-5.32679
H	3.420485	10.351006	-7.603509
H	7.127614	7.764752	-8.948617
H	7.94777	3.581009	-6.984861
H	5.082977	1.859828	-3.698812

Table F.10: Output coordinates for structure TMCPOB from the DFT calculation in COSMO solvation model. Coordinates are in atomic units.

Atom	X	Y	Z
C	3.898942	-7.12229	-2.814511
C	5.686999	-8.984907	-3.373686
C	5.657163	-11.45742	-2.368425
C	3.721803	-12.115344	-0.711768
C	1.886476	-10.316984	-0.073624
C	1.971116	-7.849167	-1.115903
C	-3.653284	-7.500515	2.910649
C	-1.461488	-8.022548	1.478813
C	-0.279837	-10.426355	1.561922
C	-1.26275	-12.361282	3.078742
C	-3.419669	-11.910434	4.516405
C	-4.54873	-9.494608	4.39473
N	-0.091951	-6.492563	-0.179716
C	4.091628	-4.562476	-4.022274
C	-5.012902	-5.011502	2.901947
C	-4.548823	-13.917781	6.173542
C	7.681162	-13.306018	-3.101252
C	-0.420785	-3.83728	-0.448084
C	-0.992214	1.425045	-0.974881
C	0.506073	0.380085	0.972328
C	0.800827	-2.203208	1.252386
C	-1.922172	-2.905913	-2.384299
C	-2.190903	-0.284558	-2.628123
C	-3.069303	4.943321	-3.633945
B	-1.351425	4.278044	-1.392297
O	-3.685759	0.469044	-4.590933
C	-4.109814	2.970857	-5.093121
C	-5.188635	7.936009	-6.448609
C	-3.661975	7.440552	-4.376875
C	-5.659503	3.431908	-7.191944
C	-6.181906	5.914836	-7.856786
C	-0.008593	6.293528	0.324805
C	3.63964	8.914927	1.266657
C	2.510661	9.794487	3.492892
C	0.102103	8.8955	4.101392
C	-1.171002	7.16355	2.555928
C	2.42046	7.186827	-0.318067
C	-3.781753	6.210862	3.282503
C	-3.711691	4.620785	5.69263
C	-5.709669	8.353292	3.453588
C	3.715561	6.239607	-2.702137
C	5.99575	4.577134	-2.083462
C	4.394385	8.388254	-4.506379
C	3.854796	11.666176	5.196249
C	6.290185	10.557957	6.288044
C	4.34129	14.17651	3.84998
H	7.203306	-8.483535	-4.688347
H	3.633989	-14.014067	0.097983
H	-0.334304	-14.206187	3.126517
H	-6.243522	-9.150113	5.529461
H	4.288492	-3.041108	-2.618326
H	2.402197	-4.114288	-5.147834
H	5.73963	-4.487343	-5.282106
H	-5.673406	-4.502302	0.997491
H	-3.796452	-3.451659	3.542915
H	-6.665147	-5.09217	4.155831
H	-3.451114	-15.678488	6.088173
H	-4.620442	-13.306275	8.162462
H	-6.503675	-14.356927	5.605988
H	7.731345	-13.597679	-5.162182
H	9.566887	-12.623256	-2.540955
H	7.382263	-15.155075	-2.204294
H	1.451413	1.674148	2.27524
H	1.9652	-3.005232	2.754865
H	-2.864331	-4.184537	-3.698545
H	-5.619532	9.880569	-6.992833
H	-2.869388	8.994065	-3.26732
H	-6.420728	1.839517	-8.261503
H	-7.38484	6.289686	-9.493603
H	5.522205	9.596922	0.752133
H	-0.800889	9.573533	5.834477
H	-4.417567	4.949769	1.743905
H	-5.597677	3.853303	6.126323
H	-3.098696	5.753799	7.328702
H	-2.39479	3.021641	5.497869
H	-7.608522	7.605973	3.863341
H	-5.217542	9.694799	4.96756
H	-5.80559	9.416102	1.667369
H	2.33953	5.029289	-3.702876
H	6.855234	3.82237	-3.823075
H	7.459275	5.652954	-1.064747
H	5.443909	2.964799	-0.888926
H	5.205626	7.640233	-6.271053
H	5.795127	9.676644	-3.662552
H	2.711992	9.510374	-4.996448
H	2.57109	12.050502	6.794517
H	7.180807	11.893275	7.613604
H	7.67347	10.144989	4.787437
H	5.91758	8.785049	7.31062
H	5.205708	15.556312	5.14712
H	5.637873	13.932188	2.238963
H	2.572383	14.993945	3.111692