doi:10.2533/chimia.2007.165 CHIMIA **2007**, *61*, No. 4

Chimia 61 (2007) 165–168

ISSN 0009–4293

# Statistical Analysis of Quantum Chemical Data Using Generalized XML/CML Archives for the Derivation of Molecular Design Rules

Andreas Elsener[§a,c,d], Claire C. M. Samson[a], Martin P. Brändle[b], Peter Bühlmann[c], and Hans P. Lüthi*[a]
[§]SCS Poster Prize Winner

*Abstract:* In this work we describe a highly automated procedure ('workflow') for the analysis of electronic and molecular structure data obtained from quantum chemical computations. The data generated as part of this workflow are archived in an XML/CML database. These data are processed by means of statistical analysis. This production and analysis machinery is applied towards the interference of dependencies between the electron delocalization and the properties of functionalized linearly π-conjugated compounds. This information is the source for the generation of rules or knowledge applicable in the rational design of functional materials.

**Keywords:** Data analysis · Electron delocalization · Electronic structure · Molecular design · Workflow

## Introduction

The rational design of novel compounds with tailored properties requires rules ('knowledge') describing the relationship between the molecular and electronic structure and the properties of the molecule. Today, this information can be generated by means of computation at very high throughput. The main challenge remaining is the transformation of this information to knowledge.

In the recent past, π-conjugated compounds have obtained considerable attention as materials for application in molecular electronics, with the vision to replace 'silicon' by 'organic materials'.[1] One important feature of π-conjugated systems is their ability to change the electronic properties in response to external stimuli (substituents, optical excitation, *etc.*). The properties of the donor/acceptor functionalized polyacetylene (PA) shown in Fig. 1 can be 'tuned' by the choice of functional groups or by the solvent used. But also the length of the PA chain (the 'backbone'), the use of spacers, *i.e.* units such as phenyl- or ethynyl-groups inserted in the oligomer chain, are additional degrees of freedom to be exploited in the design of compounds with tailored properties.

One class of compounds thoroughly investigated for their prospective use as advanced electro-optical materials are the diethynylethenes prepared and characterized in the laboratory of F. Diederich.[2] Donor/acceptor functionalized polydiacetylenes (PDA) were shown to exhibit strong electron delocalization effects, both experimentally and computationally.[3]

In this study we will focus on a well-defined set of donor/acceptor functionalized all-*trans* PDAs (Fig. 2) for which we will try to establish molecular/electronic structure–property relationships.

For the class of compounds in Fig. 2, the number of structures grows combinatorially with chain length, type and location of substituents. Each of these structures requires a set of geometry optimization, property, and electron delocalization energy calculations. The number of time-consuming manual operations involved with these calculations presents a major obstacle to high-throughput computing. Hence, an automated procedure – a workflow – which

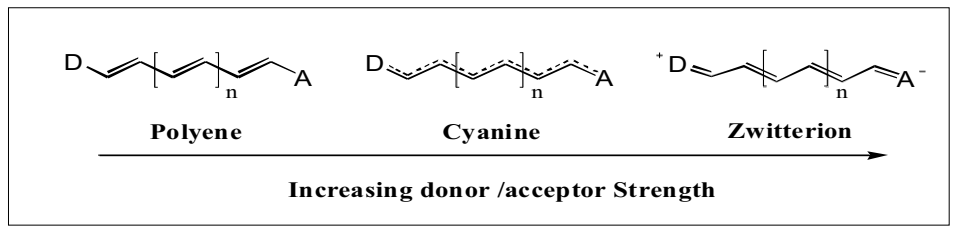


Fig. 1. Schematic representation of a functionalized π-conjugated oligomer with various degrees of bond length alternation (BLA) induced by the length of the path and the substituents

*Correspondence: PD Dr. H. P. Lüthi[a]
Tel.: +41 44 632 21 05
Fax: +41 44 632 16 15
E-Mail: luethi@phys.chem.ethz.ch
[a]Laboratory of Physical Chemistry and
[b]Chemistry Biology Pharmacy Information Center
ETH Zürich
CH-8093 Zürich
[c]Seminar for Statistics
ETH Zürich
CH-8092 Zürich
[d]Material Science and Simulation
NUM-ASQ
Paul Scherrer Institut
CH-5232 Villigen-PSI

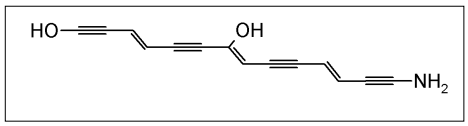


Fig. 2. Illustration of an all-*trans* PDA sample molecule (trimer with three donor substituents). An array of compounds was generated by variation of chain length (monomer to hexamer), substituent type (CN, $NO_2$, $NH_2$, and OH) and substituent position. The compound shown is referred as '1-OH-7-OH-14-NH2-all-trans-3-mer' by the workflow.

eliminates human interaction as far as possible is needed.

Inherent to workflows is the flow of data along the individual tasks. Various data formats have been suggested, of which the eXensible Markup Language (XML)[4] and the Chemical Markup Language (CML)[5] are of particular interest here because they express data as well as semantics. These markup languages are ideally suited because they allow data to be structured, exchanged, stored, extracted, and transformed. They not only carry content, but also its description, and hence are interpretable by humans and machines as well.

To establish molecular design rules from the data generated, statistical analysis was applied to find and illustrate dependencies between computed molecular properties as well as dependencies between molecular structure features and the molecular properties computed. One of the key issues in this regard was to identify dependencies between molecular properties and/or structural features and the delocalization energy. In particular, we wanted to evaluate the delocalization energy as a potential molecular descriptor. A scheme to evaluate the electron delocalization energy based on the Natural Bond Orbital (NBO) analysis had been reported in previous work (see Methods section for more detail).

In this work, we present a workflow that combines high-throughput quantum chemical calculation and statistical analysis. It consists of three parts, the quantum chemical data production, the archiving of the data in a database, and the automated statistical analysis. The technical and computational details are presented in the Methods section.

## The Workflow

### *Overview*

The workflow starts with the quantum chemical computation of the properties of each compound, including the electron delocalization energy. The data (structure, properties and meta data) are archived in a database to be retrieved for various types of statistical analysis (see Methods section for details).

Fig. 3 shows a schema of the workflow. On the left hand side is the flow of tasks, on the right hand side the flow of data. The quantum chemical part covered the following tasks:

i) input generation,
ii) geometry optimization,
iii) Natural Bond Orbital (NBO)[6] and property calculation, and
iv) the calculation of delocalization energies.

Each of the quantum chemical tasks ii–iv are dependent on the preceding task because they require data such as the equilibrium coordinates or the NBOs as their input, and therefore had to be processed sequentially.

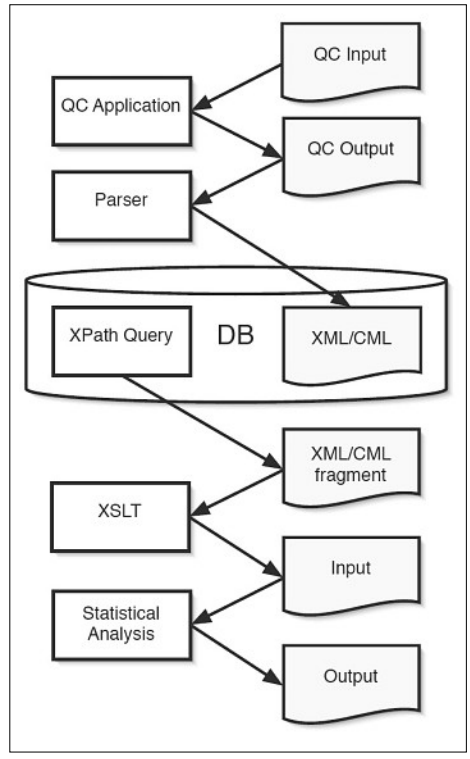


Fig. 3. Generalized workflow. On the left hand: workflow tasks; on the right hand: data flow.

Each of these tasks consisted of several steps, which included an eXtensible Stylesheet Language transformation (XSLT)[7] of the preceding XML output to produce input files, the quantum chemical calculation *per se*, and the parsing (*vide infra*) of the quantum chemical log file to produce XML/CML data. The three XML/CML outputs of tasks ii-iv were merged and transformed by XSLT to obtain the target XML/CML file, which was then archived in an XML database. One database record was produced for each molecule processed, the size of a record being between 300 and 500 KBytes. From the 1434 molecules generated in task i, 1100 could be processed completely by the workflow ii–iv. The remaining jobs failed mainly because of parsing problems. The fact that there were data from only 1100 out of 1434 compounds (full set) turned out to be irrelevant for the quality of the statistical analysis, as all subclasses were reasonably well populated with data.

### *Implementation*

The task scheduling was managed by a central scheduler implemented as an UNIX shell script. This script watched 'hot' folders for incoming files to invoke a workflow task upon their arrival. The quantum chemical calculations were performed on a separate compute cluster. For this, the necessary input files and the resulting quantum chemical outputs were transferred by Expect scripts.[8] For the queuing of the calculations, a scheduler script was implemented on the compute cluster.

The input generator was written in Java. For each compound, two files in proprietary format were generated, the quantum chemical input file and a shell script for job execution. For each compound, a name which also served as job title was generated automatically. For the nomenclature, we resorted to a plain technical description that allowed us to retrieve the data in a convenient way (see caption to Fig. 2 for an example).

### *Statistical Analysis*

Linear regression models were used in conjunction with model selection[9] to identify features that have significant influence on a response variable. Bootstrapping methods[10] were then applied to enhance the robustness of the linear models and to derive empirical probabilities for the dependencies observed. Graphical models[11] were finally used to visualize the interdependencies identified between variables (see Methods section for more detail).

We focused on two major issues associated with electron delocalization, namely on the influence of molecular and structural properties on the delocalization energy, and on the delocalization energy as a potential new molecular descriptor. To investigate the influence of molecular and structural properties on the delocalization energy, we tried to estimate empirical probabilities for the existence of dependencies. For this purpose, the model selection method for linear models was executed on 200 bootstrap samples. For predictor variables significantly used in a bootstrap sample to explain the response 'delocalization energy', the result matrix was updated accordingly with a value of 1.0. Finally we took the mean of result matrix columns to find the empirical probabilities shown in Table 1. The backbone length, the quadrupole component zz, the polarizability component yz, the ionization potential, and the total energy were used to explain the response variable 'delocalization energy' in (almost) every linear model describing a bootstrap data sample.

For the second issue, namely the electron delocalization energy being a potential

CHIMIA **2007**, *61*, No. 4

Table 1. Empirical probabilities (Prob) for dependencies between delocalization energy and structure or properties. The z-axis is perpendicular to the molecular plane.

| Data | Prob | Data | Prob | Data | Prob |
|---|---|---|---|---|---|
| End substitution 1 | 0.715 | Dipole tot | 0.025 | Polarizability zz | 0.000 |
| End substitution 2 | 0.800 | Quadrupole xx | 0.870 | Polarizability xy | 0.000 |
| Mid substitution | 0.800 | Quadrupole yy | 0.925 | Polarizability xz | 0.000 |
| Backbone length | 0.990 | Quadrupole zz | 1.000 | Polarizability yz | 1.000 |
| Mid distance | 0.000 | Quadrupole xy | 0.000 | Electron Affinity | 0.525 |
| Dipole x | 0.155 | Polarizability xx | 0.000 | Ionization Potential | 1.000 |
| Dipole y | 0.000 | Polarizability yy | 0.000 | Total Energy | 0.995 |

new molecular descriptor, we performed model selection for all the other properties. We were interested in those cases where the delocalization energy showed up as a significant predictor variable. The result for significance level 0.001 is given in Fig. 4. It shows remarkable dependencies between the delocalization energy and molecular properties.

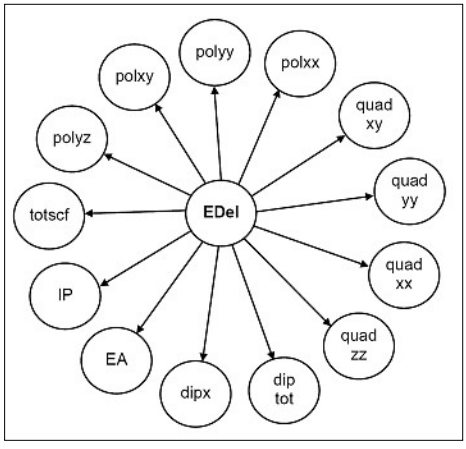


Fig. 4. Graph of properties for which the delocalization energy (EDel) is significantly used to fit a linear model (on level = 0.001). These are the quadrupoles (quad xy, quad yy, quad xx, quad zz), the dipole moment in x direction (dipx) and the total dipole moment (dip tot), the electron affinity (EA), the ionization potential (IP), the total energy (totscf), and the polarizabilities (polyz, polxy, polyy, polxx).

## Discussion and Conclusions

In the present study we show that for a given class of compounds, a large body of structural and quantum chemical information can be generated and processed in a highly automated fashion. The implementation of the data generation and processing by means of a workflow kept the number of manual intervention at a minimum, rendering a high degree of efficiency. This saved wall-clock time, and, more importantly, reduced the number of person-made errors to a minimum. It appears possible to generate and process data from thousands of compounds by means of a workflow as the one presented here.

The central part of the workflow described here consists of an XML/CML database which serves as a data repository on the one hand and as a source of data for the generation of spreadsheets for the statistical analysis on the other.

The need for *de facto* standards for electronic structure data is evident. Whereas the discipline has very high standards when it comes to the definition of model calculations, the information generated from these computations is still vastly unformatted. It appears that XML/CML is a valuable option with regard to the representation of non-binary data. Unfortunately parsers that convert proprietary application program outputs into XML/CML format are a workaround only. The ultimate solution in this respect is the definition of an XML/CML schema valid for all electronic structure application programs

The analysis of the data performed by means of linear regression and related models allowed finding dependencies between properties. Whereas some of these dependencies were expected ('must have'), others came as a surprise. An example of an unexpected relationship is the dependence between the delocalization energy and selected elements of the quadrupole tensor (Table 1). The 'must have' dependencies are important in view of the validation of the statistical methods used. It is known that the delocalization energy increases with the length of the backbone for the compounds considered here. It was also expected that a dependency between the electron delocalization and both, the ionization potential and the electron affinity (and thus the hardness/softness of the compound), would be found. The analysis clearly shows that the delocalization energy can be used to model a number of key observables (compare Fig. 4); its value as a potential new descriptor is confirmed by this study.

The chemical insight from this study is *a priori* limited due to the quality of the computed data (B3LYP/3-21G model).[12] However, with the implementation of this workflow on a Grid we will be able to perform more extensive studies on a state-of-the-art computation environment.

## Methods

### *Evaluation of the Electron Delocalization Energy*

The Natural Bond Orbital (NBO)[6] analysis is a tool to investigate chemical bonding. It also provides a measure of electron delocalization. The canonical molecular orbitals (MOs) obtained from a quantum chemical calculation, are transformed to a set of local orbitals covering lone pairs, σ- and π bonds. These natural bond orbitals (NBO) represent the Lewis structure of the molecule. The electronic charge not described by the Lewis structure, *i.e.* the delocalized charge, is hosted by anti-bonding orbitals (σ*, π*, and Rydberg).

The total binding energy of a system ($E_{tot}$) can thus be partitioned into Lewis ($E_L$) and non-Lewis ($E_{NL}$) contributions: $E_{tot} = E_L + E_{NL}$. The non-Lewis contribution ($E_{NL}$), which is a measure for the delocalization energy, is obtained by deletion of the weakly occupied π* and σ* orbitals and subsequent recomputation of the electronic energy. It has been shown that this method can successfully be applied to study electron delocalization in linearly π-conjugated compounds such as those studied here.[13,14]

### *XML for Data Flow, Archival and Retrieval*

As data format that bridges all workflow parts we choose XML[4] and its chemical representation, the chemical markup language CML.[5] For archival and retrieval of data, we used the XML native database Apache Xindice[15] which allows XML documents to be stored in collections regardless of their data structure. Data can be extracted by using Xpath[16] queries. The resulting XML outputs can be easily transformed to target formats with XSLT (eXtensible Stylesheet Language Transformation).[6,17]. The format of the XML/CML data structure used in this work can be obtained from the authors by request.

### *Statistical Analysis*

For construction of the linear models the data was transformed according to Table 2. The transformations were defined based on the analysis of scatter plots, statistical quality measures such as $R^2$ values, and statistical expert knowledge.

### *Computational Details*

Geometry optimization and the NBO analysis were carried out with Gauss-

Table 2. Variables used and their transformation for the linear regression statistical analysis.

| Variable | Transformation |
|---|---|
| Substituents | Factors |
| Backbone length | Logarithm |
| Distance of middle substitutent | Logarithm |
| Dipole moment (components and total) | Absolute value |
| Quadrupole moment (components) | Absolute value |
| Polarizability (components) | Absolute value |
| Electron Affinity (Koopman) | No transform |
| Ionization Potential (Koopman) | No transform |
| Total Energy | Logarithm of absolute value |
| Delocalization Energy | Logarithm |

ian98,[18] together with its NBO extension.[19] The calculations were performed on a Linux cluster with AMD Athlon™ XP processors. DFT/B3LYP/3-21G[12] was used, which still allows the observation of significant trends in our pioneering study. XML database queries and statistical analysis were done on a UNIX cluster Sunblade 100, Ultra Sparc IIe. The open source programs used to build and query the XML database were:

i) JUMBOMarker[20] for parsing the quantum chemistry ASCII-type output to an XML document,
ii) Saxon[21] as XSLT processor,
iii) and Xindice[15] as database.

The open source statistical analysis package R was used.[22] An important feature of this package is the option to define analysis protocols that can be executed on different data sets also in batch mode, *i.e.* rendering an easy integration of the statistical analysis in the workflow.

***Acknowledgements***

We thank Peter Murray-Rust and Joe Townsend at the Unilever Cambridge Centre for Molecular Informatics for providing us with the latest version of JUMBOMarker, the support, and the fruitful collaboration. Most of this research was performed in the context of a Master Thesis in Computational Science and Engineering at ETH Zürich.

Received: January 3, 2007

[1] a) 'Modern Acetylene Chemistry', Eds. P. J. Stang, F. Diederich, Wiley-VCH, Weinheim, **1995**; b) 'Carbon Rich Compounds II, Vol. 201', Ed. A. de Meijere, Springer, Berlin, **1999**; c) 'Conjugated Polymers and Related Materials. The Interconnection of Chemical and Electronic Structure', Eds. W. R. Salaneck, I. Lundström, B. Ränby, Oxford University Press, Oxford, **1993**; d) M. Brøndsted Nielsen, F. Diederich, *Chem. Rec.* **2002**, *2*, 189.

[2] J.-P. Gisselbrecht, N. N. P. Moonen, C. Boudon, M. B. Nielsen, F. Diederich, M. Gross, *Eur. J. Org. Chem.* **2004**, 2959–2972.

[3] M. Bruschi, PhD Thesis, ETH Zurich, No. 16343, **2005**.

[4] eXtensible Markup Language (XML), World Wide Web Consortium (W3C) *http://www.w3.org/XML/*.

[5] P. Murray-Rust, H. S. Rzepa, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 757.

[6] F. Weinhold, C. R. Landis, 'Valency and Bonding: A Natural Bond Orbital Donor-Acceptor Perspective', Cambridge University Press, **2005**.

[7] XSL Transformations (XSLT) Version 1.0, W3C Recommendation, *http://www.w3.org/TR/xslt*, **1999**.

[8] D. Libes, *Computing Systems* **1991**, *4*, 2.

[9] T. Hastie, R. Tibshirani, J. Friedman, 'The Elements of Statistical Learning: Data Mining, Inference, and Prediction' Springer, New York, **2001**.

[10] a) B. Efron, *Ann. Stat.* **1979**, *7*, 1; b) A. C. Davison, D. V. Hinkley, 'Bootstrap Methods and Their Applications', Cambridge University Press, Cambridge, **2000.**

[11] B. Frey, 'Graphical Models for Machine Learning and Digital Communication', MIT Press, Cambridge, **1998**.

[12] a) A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648; b) J. S. Binkley, J. A. Pople, W. J. Hehre, *J. Am. Chem. Soc.* **1980**, *102*, 939; c) C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785.

[13] M. Bruschi, M. G. Giuffreda, H. P. Lüthi, *Chimia* **2005**, *59*, 539.

[14] a) M. Bruschi, M. G. Giuffreda, H. P. Lüthi, *Chem. Eur. J.* **2002**, *8*, 4216; b) M. G. Giuffreda, M. Bruschi, H. P. Lüthi, *Chem. Eur. J.* **2004**, *10*, 5671.

[15] Xindice, Revision 1.1b4, The Apache Software Foundation, **2004**.

[16] XML Path Language (XPath) Version 1.0, W3C Recommendation, *http://www.w3.org/TR/xpath*, **1999**.

[17] M. Kay, 'XSLT : Programmer's Reference', 2nd ed., Wiley, Indianopolis, **2003**.

[18] Gaussian 98, Revision A.9, Gaussian, Inc., Pittsburgh PA, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, J. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, J.A. Pople, **1998**.

[19] NBO 5.0, Revision 3.1, Theoretical Chemistry Institute, University of Wisconsin, Madison, E. D. Glendening, J. K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, F. Weinhold, **2001**.

[20] JUMBOMarker, Generic Text Document Parser, Revision V 0.4, P. Murray-Rust.

[21] Saxon XSLT Processor, Revision 6.5.4, M. Key, **2005**.

[22] R: A Language and Environment for Statistical Computing, Revision 2.2.1, R Development Core Team, R Foundation for Statistical Computing, Vienna, **2006**.