Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn
Wegelerstraße 6 • 53115 Bonn • Germany
phone +49 228 73-3427 • fax +49 228 73-7527
www.ins.uni-bonn.de

M. Griebel, J. Hamaekers, F. Heber

A bond order dissection ANOVA approach for efficient electronic structure calculations

INS Preprint No. 1402

March 2014

A bond order dissection ANOVA approach for efficient electronic structure calculations

Michael Griebel, Jan Hamaekers, and Frederik Heber

Abstract In this article, we present a new decomposition approach for the efficient approximate calculation of the electronic structure problem for molecules. It is based on a dimension-wise decomposition of the space the underlying Schrödinger equation lives in, i.e. $\mathbb{R}^{3(M+N)}$, where M is the number of nuclei and N is the number of electrons. This decomposition is similar to the ANOVA-approach (analysis of variance) which is well-known in statistics. It represents the energy as a finite sum of contributions which depend on the positions of single nuclei, of pairs of nuclei, of triples of nuclei, and so on. Under the assumption of locality of electronic wave functions, the higher order terms in this expansion decay rapidly and may therefore be omitted. Furthermore, additional terms are eliminated according to the bonding structure of the molecule. This way, only the calculation of the electronic structure of local parts, i.e. small subsystems of the overall system, is necessary to approximate the total ground state energy. To determine the required subsystems, we employ molecular graph theory combined with molecular bonding knowledge. In principle, the local electronic subproblems may be approximately evaluated with whatever technique is appropriate, e.g. HF, CC, CI, or DFT. From these local energies, the total energy of the overall system is then approximately put together in a telescoping sum like fashion. Thus, if the size of the local subproblems is independent of the size of the overall molecular system, linear scaling is directly obtained. We discuss the details of our new approach and apply it to both, various small test systems and interferon alpha as an example of a large biomolecule.

Michael Griebel
Institute for Numerical Simulation, University of Bonn, Wegelerstr. 6, 53115 Bonn, Germany, e-mail: griebel@ins.uni-bonn.de

Jan Hamaekers
Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, 53754 Sankt Augustin, Germany, e-mail: jan.hamaekers@scai.fraunhofer.de

Frederik Heber
Institute for Numerical Simulation, University of Bonn, Wegelerstr. 6, 53115 Bonn, Germany, e-mail: heber@ins.uni-bonn.de

1 Introduction

The coupling of the micro- and the mesoscale of chemical reactions is currently a field of intensive research. Where the microscale is the realm of quantum mechanical effects, the mesoscale is described by statistical mechanics and macroscopic thermodynamics. Nevertheless, there are additional strong influences onto the mesoscale by effects from the microscale. Numerically, the microscale is usually treated with Hartree-Fock (HF), Configuration Interaction (CI), Coupled Cluster (CC), or Density Functional Theory (DFT) methods which yield approximate solutions to the underlying quantum-mechanical (QM) Schrödinger equation (SE), whereas the mesoscale is covered by classical molecular mechanics (MM) methods which use Newton's mechanics with empirically fitted potential functions.

The ultimate goal would be a seamless coupling of quantum mechanical computations where needed and classical molecular mechanics simulations where sufficient. Such approaches are generally referred to as multi-scale methods, an extensive overview is given in [28]. Any starting point must be the general Schrödinger equation for the electrons and nuclei of the system under consideration. The Schrödinger equation however lives in $3(M+N)$ dimensions, where M denotes the number of nuclei and N denotes the number of electrons. This renders a direct numerical treatment impossible due to the curse of dimension and one has to resort to model approximations. As a first step, in the Born-Oppenheimer molecular dynamics (MD) approach, the wave functions of the nuclei and electrons are separated, the subsystem of the nuclei is treated classically with Newton's mechanics and the remaining $3N$-dimensional electronic Schrödinger equation is further approximated by one of the aforementioned methods. The potential needed for Newton's mechanics is obtained from the electronic solution by the Hellmann-Feynman theorem. This way, QM and MM are globally coupled. However, a global electronic QM solution is, at least for larger molecules, still too expensive as conventional methods scale at best with $O(M^3)$ due to the underlying problem of matrix diagonalization.

Thus, general *linear scaling* electronic structure methods are employed to overcome the dimensionality problem. As a first step, for the long-range Coulomb interaction, the use of the fast multipole method [16] has resulted in $O(M \log M)$ complexity. Furthermore, a cutoff radius such as for the MP2 theory [5] was used in a Divide&Conquer approach to take advantage of the exponential decay properties of electronic wave functions. Altogether, this resulted in linear scaling [14]. Another common method is the Density Matrix Minimization technique [27, 8]. There, the density matrix is unconstrainedly minimized via a conjugate gradient scheme, using idempotency and normalization. The Fock matrix is the minimized output, after off-diagonal elements have also been truncated at a cutoff radius. The electronic localization in non-metallic systems can also be exploited for plane wave basis sets [34]. Again, a cutoff then allows for linear scaling. Note however that there is a crossover point up to which the standard cubic scaling approaches still perform faster due to smaller prefactors in their computational complexity counts [15].

In order to reduce the constants and thus to shift this crossover point, one tries to somehow further decompose the full global electronic Hamiltonian into local parts

and employs local QM there. Let us briefly summarize the most common *decomposition approaches* in the following. One of the first, the Force-Matching Method by Ercolessi [13], tries to automatically generate empirical potentials by a least-square fitting of the forces of ab-initio calculations to general many-body potential forms such as that of the Embedded Atom approach [11] or that of Abell-Tersoff [1, 36]. Then, there is a range of methods which employ a decomposition[1] directly in $\mathbb{R}^3$, like e.g. the SIBFA (Sum of Interactions Between Fragments computed Ab initio) procedure [17] and its generalization, the so-called Fragmentation Reconstruction Method (FRM) [2]. Further schemes are the so-called IMOMM ansatz proposed by Morokuma [30], the ONIOM approach [39] and the well-known Fragment Molecular Orbital (FMO) method [25]. A scheme for modeling the electrostatic impact of a passive MM environment on the active QM system is described in [26]. Moreover, in [3] and [33] an interface regime between QM and MM with "link" atoms is proposed to account for the cutting of bonds. Similar techniques are used in [37, 38]. The common basic idea of most approaches is to use a telescoping sum over two regions to describe the total energy, like e.g. Ω_1 and $\Omega_2 \subseteq \Omega_1$, where the energy is split as $E^{QM/MM} = E^{MM}_{\Omega_1} + E^{QM}_{\Omega_2} - E^{MM}_{\Omega_2}$, where to our knowledge all procedures involve stringent chemical knowledge to choose the regions (or cuts) as best as possible but to still keep the ground-state electronic density intact. Another approach divides *time* instead of space in order to generate a coupling between QM and MM. One of these methods is learn-on-the-fly [10], which is similar to the Force-Matching method, but is run during the computation: At intermittent time steps certain clusters of the simulation domain are locally computed by QM, and the obtained local forces are used to correct the MM calculation.

While all of the above methods have promising features, we feel that they generally either involve too many additional parameters, unchemically cut bonds in separating active from passive regions, or even worse, add unphysical pseudo-atoms in order to compensate for the different energy and time scales and to avoid spill-out effects of electronic density or energy. Moreover, they are plainly too simple or do not grasp the problem in its full complexity, since only a matching or interpolation with respect to energy or forces between the QM and MM parts of the overall approach is employed.

There is one more group of methods that build upon *additivity models*, well-known in chemistry, see [19] and references therein. The central idea is to construct molecular properties of a system by adding up the corresponding known properties of its fragments. The principal hope is that a high-dimensional system such as a complex molecule depends strongly only on few input variables. Rabitz et al [19] describe a High-Dimensional Model Representation (HDMR) that can also be understood as an *ANalysis Of VAriance* (ANOVA), which is well-known from statistics. They address the problem of the estimation of the enthalpy of formation of a broad range of organic molecules based on experimental data, but they do not assess the possibilities of the ansatz in the field of electronic structure calculations. Deev and Collins [12, 9] use this additivity model ansatz by calculating the total electronic

[1] Ultimately, the aim would be a decomposition of $\mathbb{R}^{3(M+N)}$, the space where the full Schrödinger equation lives in.

energy of fragments of a system under consideration to obtain a good approximation of the energy of the total molecule. They do give an algorithmic description, however which we feel is not fully consistent with the mathematical basics, governed by the ANOVA or HDMR scheme.

In this article, we propose a more sophisticated algorithm. The additivity models place their hope on the same grounds as do many-body potential such as Tersoff's [36], where the energy and the forces of an atom are assumed to depend on its local coordination. Here, for a proof-of-concept, we concentrate on covalent bonding, hence on charge-neutral molecular systems and subsystems.[2] We will use this knowledge of coordination and bonds between nuclei to decompose the space $R^{3(M+N)}$ of the underlying Schrödinger equation in a dimension-wise fashion. This decomposition is similar to the ANOVA-approach. It represents the energy as a finite sum of contributions which depend on the positions of single nuclei, of pairs of nuclei, of triples of nuclei, and so on. Under the assumption of locality of electronic wave functions, the higher order terms in this expansion decay rapidly and may therefore be omitted. Furthermore, additional terms are eliminated according to the bonding structure of the molecule. This way, only the calculation of the electronic structure of local parts, i.e. small overlapping subsystems of the overall molecule system, is necessary to approximate the total ground state energy. To determine the required subsystems, we employ molecular graph theory combined with molecular bonding knowledge. Here, modern graph algorithms are used to create proper local subproblems as overlapping fragments of the overall molecular system. Furthermore, hydrogenization is used to close shells and saturate bonds that have been cut. We thus also exploit locality, however not by an explicit cutoff radius as most conventional methods do, but by implicitly using it in the inherent bond structure of the molecular system. In principle, the local electronic subproblems may be approximately evaluated with whatever QM technique is appropriate, e.g. HF, CI, CC, or DFT. From these local energies, the total energy of the overall system is then approximately put together in a telescoping sum like fashion. Thus, if the size of the local subproblems is independent of the size of the overall molecular system, linear scaling is directly obtained. The $3(M+N)$-dimensional full global Hamiltonian is broken down within the Born-Oppenheimer Approximation to $O(M)$ components, the i-th of them with $M_i^{(k)}$ degrees of freedom, with an upper bound $\max_i\{M_i^{(k)}\}$ controlled by a single parameter k which we name the *bond order* of the approximation. This ansatz specifically combines the smaller prefactor of the cubic scaling methods with a general linear scaling behavior. As the size of each subproblem depends on the bond coordination of the involved atoms, we coined the method BOSSANOVA (Bond Order diSSection ANOVA).

The remainder of this article is organized as follows: In Sect. 2 we briefly summarize the basics of the underlying Schrödinger equation. In Sect. 3 we describe the ANOVA-like decomposition of the energy of the Schrödinger equation in the context of molecular graph theory. In Sect. 4 we give numerical results for a broad range of organic molecules. We end with some concluding remarks in Sect. 5.

[2] Note however that our approach should work equally well also in the non-charge neutral case.

2 Schrödinger Equation in the Born-Oppenheimer Approximation

Let us consider a molecular system consisting of M nuclei and N electrons. Its time-dependent state function can be written in general as

$$\Psi = \Psi(R_1, \ldots, R_M, r_1, \ldots, r_N, t),$$

where R_i and r_j denote positions in three-dimensional space $\mathbb{R}^3$ associated to the ith nucleus and the jth electron, respectively. The variable t denotes the time-dependency of the state function. The vector space (space of configurations), in which the coordinates of the particles are given, is therefore of dimension $3(M+N)$. In the following we will abbreviate $(R_1, \ldots, R_M)$ and $(r_1, \ldots, r_N)$ with the shorter notation $\mathbf{R}$ and $\mathbf{r}$, respectively. Also, we assume that Ψ is normalized to $\int \Psi^*(\mathbf{R},\mathbf{r},t)\Psi(\mathbf{R},\mathbf{r},t)d\mathbf{R}d\mathbf{r} = 1$.

Nuclei and electrons are charged particles. The electrostatic potential (Coulomb potential) of a point charge is $1/r$ in atomic units, where r is the distance from the position of the charged particle. An electron moving in this potential possesses the potential energy $V(r) = -1/r$. Neglecting spin and relativistic interactions and assuming that no external forces act on the system, the Hamilton operator in position representation associated to the system of nuclei and electrons is given as the sum over the operators for the kinetic energy and the Coulomb potentials,

$$\begin{aligned} &H(N,M,Z_1,m_1,\ldots,Z_M,m_M;\mathbf{R},\mathbf{r}) := \\ &\underbrace{-\frac{1}{2}\sum_{k=1}^{N}\Delta_{r_k} + \sum_{k<j}^{N}\frac{1}{\|r_k - r_j\|} - \sum_{k=1}^{N}\sum_{j=1}^{M}\frac{Z_j}{\|r_k - R_j\|} + \sum_{k<j}^{M}\frac{Z_k Z_j}{\|R_k - R_j\|}}_{H_e(N,M,Z_1,m_1,\ldots,Z_M,m_M;\mathbf{R},\mathbf{r})} - \frac{1}{2}\sum_{k=1}^{M}\frac{1}{m_k}\Delta_{R_k}, \end{aligned} \tag{1}$$

where we use a semicolon to distinguish between parameters (i.e. the number M of atoms, the number N of electrons, the nuclei mass in atomic units m_j and the atomic number Z_j) and the degrees of freedom (i.e. the positions $\mathbf{R}$ and $\mathbf{r}$). Here, $\|r_k - r_j\|$ are the distances between electrons, $\|r_k - R_j\|$ are distances between electrons and nuclei and $\|R_k - R_j\|$ are distances between nuclei. We will omit parameters from this list if they are clear from the context. This will later especially be $N,M,Z_1,m_1,\ldots,Z_M,m_M$.

Now, a system of equations for the electronic and for the nuclei degrees of freedom is usually derived with the *Born-Oppenheimer approximation*. To this end, the large difference in masses between electrons and atomic nuclei is exploited to decouple the motion of the electrons from that of the nuclei.[3] Then, one assumes that

[3] The ratio of the velocity v_K of a nucleus to the velocity of an electron v_e is in general smaller than 10^{-2}.

the electrons adapt instantaneously to a change in the nuclear configuration and are thus always in the quantum mechanical ground state denoted by $\phi_0(\mathbf{R}(t);\mathbf{r})$, which is associated to the actual position of the nuclei $\mathbf{R}(t)$. Note that this allows us to write $H_e(\mathbf{R}(t);\mathbf{r})$ instead of $H_e(\mathbf{R}(t),\mathbf{r})$ since the movement of the nuclei during the adaptation of the electron positions is negligibly small in the sense of classical dynamics. This justifies to set $\Psi(\mathbf{R},\mathbf{r},t) \approx \Psi^{BO}(\mathbf{R},\mathbf{r},t) := \sum_{j=0}^{\infty} \chi_j(\mathbf{R},t)\phi_j(\mathbf{R};\mathbf{r})$, which allows to separate the fast from the slow variables. We then obtain the following set of equations:

$$M_k \ddot{R}_k(t) = -\nabla_{R_k} \underbrace{\min_{|\phi_0(\mathbf{R}(t);\cdot)|=1} \left\{ \int \phi_0^*(\mathbf{R}(t);\mathbf{r}) H_e(\mathbf{R}(t);\mathbf{r}) \phi_0(\mathbf{R}(t);\mathbf{r}) d\mathbf{r} \right\}}_{=: V_e^{BO}(\mathbf{R}(t))} \tag{2}$$

$$H_e(\mathbf{R}(t);\mathbf{r})\phi_0(\mathbf{R}(t);\mathbf{r}) = E_0(\mathbf{R}(t))\phi_0(\mathbf{R}(t);\mathbf{r}). \tag{3}$$

In the end, after time discretization, we have to perform in each time step the following tasks: First, we have to compute an approximate solution of the electronic Schrödinger equation in (3) for fixed positions $\mathbf{R}$ of the nuclei, then we have to compute from its solution the forces on the nuclei and finally we have to compute the positions of the nuclei at the next time step by e.g. a Verlet time step for Newton's equations of motion of the nuclei in (2). To this end, we use the *Hellmann-Feynman Theorem* to obtain the electronic forces

$$F_k(\mathbf{R}(t)) = -\nabla_{R_k} \int \phi_0^*(\mathbf{R}(t)) H_e(\mathbf{R}(t)) \phi_0(\mathbf{R}(t)) d\mathbf{r}$$

acting on the nuclei. Variants of this approach are the Ehrenfest molecular dynamics and the Car-Parrinello method. For details of the derivation, see [18] and the references cited therein.

3 ANOVA Decomposition Scheme

So far, the Born-Oppenheimer molecular dynamics was employed to split the full Schrödinger problem into two parts, i.e. a classical Newton's equation of motion for the nuclei, and, in each discretized time step, the electronic eigenproblem (3) which may approximately be solved by e. g. the Hartree Fock, Configuration Interaction, Coupled Cluster, or Density Functional method, see [35, 31]. However, such an overall approach is only feasible for small molecules due to the high complexity of any approximate solution method for the electronic problem. To overcome this difficulty, the aforementioned coupling techniques and linear scaling methods had

been developed. They basically all exploit locality of the electronic wave function in one way or another to reduce the complexity of the electronic problem.[4] [5]

In the following, we also resort to a certain locality of the electronic wave function. It is expressed in the bond structure of the molecular system. We decompose the overall electronic problem into small subproblems which then may be handled efficiently. To this end, we introduce an ANOVA decomposition scheme for the energy of a molecular system into local parts by means of the bond order of the nuclei in the system.

3.1 ANOVA Expansion

We will now define the energy function for a molecular system and its ANOVA series expansion. To this end, we consider a molecular system which consists of N electrons and M nuclei, each with coordinate vector $R_i \in \mathbb{R}^3$ and atomic number $Z_i \in \mathbb{N}$, $i \in \{1,\dots,M\}$. We restrict ourselves to charge-neutral systems, i.e. the number of electrons N is equal to $\sum_i^M Z_i$ for reasons of simplicity. Finally, we consider the systems only in their electronic ground state in the framework of the Born-Oppenheimer molecular dynamics. To this end, we separate the time-independent electronic Schrödinger equation as in (3) and define a total ground state energy function $E^M : (\mathbb{N}\times\mathbb{R}^3)^M \to \mathbb{R}$. It depends on the parameters that completely identify the system under consideration, namely the coordinates R_i and the atomic number Z_i of each nuclei with fixed and unique label $i \in \{1,\dots,M\}$, i.e.

$$E^M(\underbrace{(Z_1,R_1)}_{=:X_1},\dots,\underbrace{(Z_M,R_M)}_{=:X_M}) := \min_{|\phi_0(\mathbf{R}(t);\cdot)|=1} \int \phi_0^*(\mathbf{R}(t);\mathbf{r}) H_e\big(N=\sum_{i=1}^M Z_i, X_1,\dots,X_M\big)\phi_0(\mathbf{R}(t);\mathbf{r})\big)\,d\mathbf{r}, \quad (4)$$

where we further simplify the notation by defining $X_i := (Z_i,R_i)$. I. e. X_i combines the atomic number and the coordinates of the nuclei i. Note that, due to the charge-neutrality condition $N=\sum_i^M Z_i$, the parameter N may now be eliminated from the parameter list of the Hamiltonian H.

Now we will decompose the function E^M in a multivariate telescoping sum, i.e. in a finite series expansion in the nucleic parameters, in a similar way as the ANOVA decomposition[6] [21]. This decomposition involves a splitting of the M-dimensional

[4] This excludes in general metallic systems, whose electrons may be delocalized due to a vanishing band gap.

[5] Furthermore, the notion of the locality of the wave function is important as it leads to the general chemical understanding of molecules from the general bond structure up to nucleophilic sites.

[6] The ANOVA decomposition of a M-dimensional function $f:[0,1]^M \to \mathbb{R}$ reads $f=\sum_{u\subseteq\{1,\dots,M\}} f_u$ with f_u depending only on the variables indicated in u. The functions f_u satisfy the recurrence relation $f_\emptyset = L_{\{1,\dots,M\}}(f)$, $f_u = L_{\{1,\dots,M\}/u}(f) - \sum_{v\subset u} f_v$ with $L_w(f) = \int_{[0,1]^{|w|}} f(x_1,\dots,x_M)\,dx_w$. Thus,

function into contributions which depend on the positions of single nuclei and associated charges, of pairs of nuclei and associated charges, of triples of nuclei and charges, and so on. To this end, we consider the subset of the nuclei parameters $\{X_i\}_{i\in I}$ described by a set of labels I with cardinality $|I| = k$ and call it the *molecular fragment* associated to I with size k. Note that we here do not need to consider the electronic degrees of freedom $\mathbf{r}$, as the system is assumed to be in ground state and, hence, the electronic state functions are all fixed by the minimum condition in (4).

First, we define the total electronic ground state energy of lower-dimensional subsystems of the molecular system under consideration, described by the set of indices $I = \{i_1, \ldots, i_k\}$,

$$E_{\{i_1,\ldots,i_k\}}(X_1,\ldots,X_k) := \min_{|\phi_0|=1} \int \phi_0^*(\mathbf{r}) H_e\big(\sum_{j=1}^{k} Z_{i_j}, X_{i_1},\ldots,X_{i_k}\big)\phi_0(\mathbf{r})\big)d\mathbf{r}. \qquad (5)$$

Note that this is in form very similar to (4). In the notation of the electronic ground state wave functions ϕ_0, the dependency on $\mathbf{R}(t)$ was dropped as it is clear from the context.

Then, the energy function E^M is decomposed analogously to the ANOVA approach as

$$\begin{aligned} E^M(X_1,\ldots,X_M) = {} & F_\emptyset \\ & + \sum_{i_1}^{M} F_{\{i_1\}}(X_{\{i_1\}}) \\ & + \sum_{i_1<i_2}^{M} F_{\{i_1,i_2\}}(X_{\{i_1,i_2\}}) \\ & + \sum_{i_1<i_2<i_3}^{M} F_{\{i_1,i_2,i_3\}}(X_{\{i_1,i_2,i_3\}}) \\ & + \ldots \\ & + F_{\{i_1,\ldots,i_M\}}(X_{\{i_1,\ldots,i_M\}}) \\ & =: \sum_{U\subseteq\{1,\ldots,M\}} F_U(X_U), \end{aligned}$$

where X_U denotes the set of variables $\{X_i\}_{i\in U}$ and $U \subseteq \{1,\ldots,M\}$.

Here, each term $F_{\{i_1,\ldots,i_k\}}$ is defined as follows:

---

f is decomposed into a constant, a sum of one-dimensional functions, a sum of two-dimensional functions, and so on. The involved functions are generated by proper partial integration and telescopic corrections according to the recurrence relation.

$$
\begin{aligned}
F_{\emptyset} &= 0 \\
F_{\{i_1\}}(X_{\{i_1\}}) &= \gamma_{\{i_1\}}\big(E_{\{i_1\}}(X_{\{i_1\}}) - F_{\emptyset}\big) \\
F_{\{i_1,i_2\}}(X_{\{i_1,i_2\}}) &= \gamma_{\{i_1,i_2\}}\big(E_{\{i_1,i_2\}}(X_{\{i_1,i_2\}}) - F_{\{i_1\}}(X_{\{i_1\}}) - F_{\{i_2\}}(X_{\{i_2\}}) - F_{\emptyset}\big) \\
&\dots \quad \dots \\
F_{\{i_1,\dots,i_k\}}(X_{\{i_1,\dots,i_k\}}) &= \gamma_{\{i_1,\dots,i_k\}}\big(E_{\{i_1,\dots,i_k\}}(X_{\{i_1,\dots,i_k\}}) \\
&\quad - \sum_{U\subseteq I,|U|=k-1} F_U(X_U) \\
&\quad - \sum_{U\subseteq I,|U|=k-2} F_U(X_U) \\
&\quad \dots \\
&\quad - \sum_{U\subseteq I,|U|=1} F_U(X_U) - F_{\emptyset}\big) \\
&\dots \quad \dots,
\end{aligned}
$$

where the constant function $F_{\emptyset}$ is set equal to zero since it corresponds to the energy of an empty molecular system and a set $\{\gamma_I\}_{I\subseteq\{1,\dots,M\}}$ of weights $\gamma_I \in \{0,1\}$ is involved to switch on and off the considered interaction terms. I.e. we have

$$E^M(X_1,\dots,X_M) = \sum_{U\subseteq\{1,\dots,M\}} F_U(X_U), \tag{6}$$

where

$$F_U(X_U) = \gamma_U\big(E_U(X_U) - \sum_{k=0}^{|U|-1} \sum_{V\subseteq U,|V|=k} F_V(X_V)\big) \tag{7}$$

and $E_{\emptyset} = 0$. Let us for the moment assume that all γ_I are set to one. Then the decomposition is exact and contains 2^M different terms due to the power set construction. In general it might be that all terms are equally important up to the last, M-dimensional one, or, in the extreme case that the last term might be the only important one and thus nothing is gained from this decomposition. However, if the size of the terms decay fast with e.g. the order of the terms, then a proper truncation of the ANOVA series expansion results in a substantial reduction in computational complexity. We then only have to deal with a sequence of lower-dimensional subproblems which are associated to the remaining lower-dimensional energy terms of the decomposition.

Let us remark that the energy functions $F_{\{i_1,\dots,i_k\}}$ in (6) may be recognized as an expansion of many-body interaction contributions, as in [29]. This leads us to the following assumption which is central to our further approach: There is a certain decay in the contribution of each order k of the ANOVA expansion and this results in a monotone convergence of the approximation error with rising order. Consequently, from a certain order onward, we may neglect the higher higher-order terms in the

ANOVA decomposition. This results in a good approximation to the true result[7] with an accuracy which is related to the order parameter at which the truncation was executed. This assumption is also strongly supported by the success of conventional two- and many-body potential functions used in classical molecular dynamics, such as short range pair-potentials like harmonic springs, the Morse potential and the Lennard-Jones potential, three- and four-body potential like angle and dihedral potential functions and more advanced many-body potential functions which involve a local coordination number (that is the local density of atoms) like Tersoff's potential [36], the embedded atom method [11] or Brenner's reactive bond order potential for hydrocarbons [7]. Here, in any case, only a small number of neighboring atoms are involved in the potential forms, for further details see [18].

Our ansatz is as follows: We decompose the total energy function (4) in an ANOVA series expansion as in (6) where we include only terms up to a certain order k, which we call the *bond order* of the approximation. Now, let $G = (P,K)$ be the associated graph that represents the bond structure of the molecular system under consideration. For reasons of simplicity we assume that this graph is connected. Then, we neglect in a second step even further interaction terms in the ANOVA expansion. These terms contain as parameters the degrees of freedom which belong to nuclei in I that are not connected by a path in the graph G_I, i. e. we additionally eliminate those terms whose induced subgraph G_I is not connected by setting γ_I to zero. Note here that each set $I = \{i_1, \ldots, i_{|I|}\}$ of nuclei parameters indices for each term $E_{\{i_1,\ldots,i_{|I|}\}}(X_{\{i_1,\ldots,i_{|I|}\}})$ in (6) is directly associated to an induced subgraph $G_I = (P_I, K_I)$ of the total graph G with $P_I = \{v_i\}_{i\in I}$ and $K_I = \{\{v_1, v_2\} \in K : v_1 \in I, v_2 \in I\}$. This second elimination step is motivated by the locality of the electronic wave functions: Atoms that share a bond to a nearby atom will be strongly influenced by changes in the chemical vicinity of nearest or next-nearest bonding partners whereas atoms that share no bond to a nearby atom will not.

Altogether, this can be described by an approximation to the ground state energy according to (6) and (7). To this end, let $G = (P,K)$ be the interaction graph of the molecular system under consideration. We then define a set of graph-related weights $\{\gamma_U^G\}_{U\subseteq\{1,\ldots,M\}}$ by

$$\gamma_U^G = \begin{cases} 1, & \text{if the subgraph } G_U \text{ of } G \text{ (induced by } U\text{) is connected,} \\ 0, & \text{else.} \end{cases} \tag{8}$$

This definition is motivated from the following observation. Let us assume that $E_{A\bigcup B}(X_A, X_B) = E_A(X_A) + E_B(X_B)$ for all pairs of disconnected subgraphs G_A and G_B which are induced by disjoint subsets $A, B \subseteq P$, $A \bigcap B = \emptyset$ and for simplicity let us further assume that all weights are set to one. Then we can derive the following statement:

[7] Note that, in practice, the global electronic problem is only solved approximately anyway, by e.g. DFT, CC, CI.

Lemma 1. *Let $G=(P,K)$ be an interaction graph. Let $A,B\subseteq P$, $A\bigcap B=\emptyset$ and let the subgraphs G_A and G_B induced by A and B, respectively, be disconnected. Then*

$$F_{A\bigcup B}(X_{A\bigcup B})=0.$$

Proof. We use induction: The base case can be easily seen for graphs $G=(P,K)$ with sets $|P|\leq 2$. Let us assume that the statement holds for graphs $G=(P',K')$ with $|P'|\leq n$. Now let $G=(P,K)$ with $|P|=n+1$. Note that from the recursive definition of F_U it immediately follows that

$$E_U(X_U)=\sum_{u\subseteq U}F_U(X_U)$$

holds for all $U\subseteq P$. With the assumption $E_{A\bigcup B}(X_{A\bigcup B})=E_A(X_A)+E_B(X_B)$ and $F_\emptyset=0$, we then obtain

$$\begin{aligned}F_{A\bigcup B}(X_{A\bigcup B}) &= E_{A\bigcup B}(X_{A\bigcup B})-\sum_{a\subseteq A,a\neq\emptyset}F_a(X_a)-\sum_{b\subseteq B,b\neq\emptyset}F_b(X_b)\\ &\quad-\sum_{\substack{a\subset A,b\subset B\\ a\neq\emptyset,b\neq\emptyset,|a\bigcup b|<|A\bigcup B|}}F_{a\bigcup b}(X_{a\bigcup b})-F_\emptyset\\ &= E_A(X_A)+E_B(X_B)\\ &\quad-\sum_{a\subseteq A}F_a(X_a)-\sum_{b\subseteq B}F_b(X_b)-\sum_{\substack{a\subset A,b\subset B\\ a\neq\emptyset,b\neq\emptyset,|a\bigcup b|<|A\bigcup B|}}F_{a\bigcup b}(X_{a\bigcup b}).\end{aligned}$$

Now, we apply the induction hypothesis to each $F_{A\bigcup B}$: $|a\bigcup b|<|A\bigcup B|\leq|P|=n+1$ and finally obtain

$$\begin{aligned}F_{A\bigcup B}(X_{A\bigcup B}) &= E_A(X_A)-\sum_{a\subseteq A}F_a(X_a)+E_B(X_B)-\sum_{b\subseteq B}F_b(X_b)\\ &\quad-\sum_{\substack{a\subset A,b\subset B\\ a\neq\emptyset,b\neq\emptyset,|a\bigcup b|<|A\bigcup B|}}F_{a\bigcup b}(X_{a\bigcup b})\\ &= -\sum_{\substack{a\subset A,b\subset B\\ a\neq\emptyset,b\neq\emptyset,|a\bigcup b|<|A\bigcup B|}}F_{a\bigcup b}(X_{a\bigcup b})=0.\end{aligned}$$

□

3.2 Saturation with Hydrogen

After the motivation of the basic principles of our decomposition scheme in the last section, we now have to face a technical difficulty: A cut-out fragment may have a total spin unequal zero while the molecular system itself has a total spin of zero.

As closed-shell calculations are algorithmically both simpler and more stable, this situation would complicate the proposed linear-scaling ansatz.

A step to remedy this situation is a saturation of the dangling bonds of the fragments by adding hydrogen at the places where bonds were cut, causing the total spin of the fragment system to be zero. Due to our telescopic sum approach the effect of the hydrogen atoms actually goes unnoticed.

This correction is schematically depicted in Fig. 1 where we just show two atoms and its vertex but omitted for simplicity any further vertices and edges these atoms might be connected to. Here, let us assume that, after cutting the edge k_i, Atom1 should belong to an induced subgraph G', while Atom2 should not. Then, edge $k_i = \{\text{Atom1, Atom2}\}$ is not present in this subgraph. Now, we insert two new terminal vertices H1 and H2 and two new edges $k_1^{(H)} = \{\text{Atom1, H1}\}$ and $k_2^{(H)} = \{\text{Atom2, H2}\}$ so that all dangling bonds are closed. Hence, the new vertex H1 and the edge $k_1^{(H)}$ would be added to G' next to Atom1. By this saturation procedure, we only calculate closed-shell atoms. In particular, the electronic density of the cut edges is thus conserved to a higher degree. Note that this approach is still tunable by the bond length used between new hydrogen vertices and cut-vertices. In our subsequent implementation we use here the equilibrium hydrogen bond lengths of certain small molecules taken from [24].

(a) Edge k_i before cutting (b) Edge k_i disconnected

Fig. 1 Cut of an edge k_i between two vertices and replacement with two edges $k_1^{(H)}$ and $k_2^{(H)}$ to two newly introduced terminal vertices (hydrogen atoms H1 and H2).

This saturation procedure can be understood as a re-definition of the electronic Hamiltonian H_e in (5): From the known graph G of the molecular system l additional hydrogen vertices, bonds and their graph-dependent coordination $R_i^H(G)$, $1 \leq i \leq l$, are derived and the ground state energy evaluated for this system is defined as:

$$\widehat{E}_{i_1,\ldots,i_k}(X_1,\ldots,X_k) := \min_{|\phi_0|=1} \int \phi_0^*(\mathbf{r}) H_e\big(l + \sum_{j=1}^{k} Z_{i_j}, X_{i_1},\ldots,X_{i_k}, R_1^H(G),\ldots,R_l^H(G)\big)\phi_0(\mathbf{r})d\mathbf{r}. \quad (9)$$

Note that this saturated energy function is denoted by $\widehat{E}$.

The saturation procedure by means of hydrogen renders the role of hydrogen special in our approach. Thus, it is is useless to cut out a fragment at an edge involving only one hydrogen nucleus, as this will only create an additional hydrogen molecule while leaving the edge as it was before. Here, the best procedure is to remove the hydrogen nuclei degrees of freedom from the ANOVA decomposition

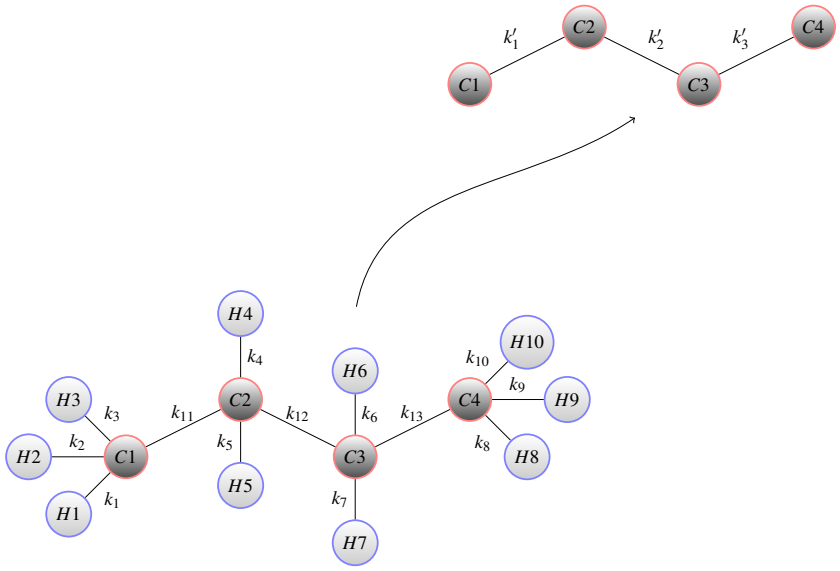


Fig. 2 Hydrogen vertices in light gray are combined with their bonding partners in dark gray to new single vertices. The remaining edges and new vertices have been relabeled, denoted by single digits.

algorithm, i. e. to drop them completely from the graph G, or to combine them with their bonding partners since they are always terminal vertices anyway, see Fig. 2 for an illustration. Hence, in the following, we will not take further heed of the hydrogen atoms which are present in the molecular system. This is also advantageous since e.g. about half of the atoms in organic molecules are hydrogens. Thus, we strongly reduce the necessary number of fragments to be evaluated.

Altogether, to a given bond graph G we define the BOSSANOVA approximate energy up to order k by

$$E^{\mathrm{ANOVA}}(k) = \sum_{U \subseteq \{1,\dots,M\}, |U| \leq k} F_U(X_U), \tag{10}$$

with F_U according to the recursive definition (7) using energies determined by (9) and weights $\{\gamma_U^G\}_{U \subseteq \{1,\dots,M\}}$ chosen like in (8).

3.3 Scaling Behavior

We give some theoretical limits on the scaling behavior of the proposed approach along with a constructive proof which our actual implementation follows closely. Here, just the dependence of the number of fragments is to be considered.

The maximum number of fragments possible for a molecular system consisting of M nuclei is given by the power set 2^M. Generally, we obtain for the power set, truncated to contain at most k nuclei, the following relation $\sum_{l=0}^{k} \frac{M!}{l!(M-l)!} \approx M^k$ for small k. However, in our ansatz many fragments are discarded when they do not constitute a connected subgraph of the molecular system. Hence, the true number of fragments is actually a lot smaller as is shown with the following lemma.

Lemma 2 (Upper bound on number of connected subgraphs). *Let a connected graph $G = (P,K)$ be given. Let the number of edges per vertex be bounded from above by $c > 0$.*

Then, the number of induced subgraphs $G' = (P',K')$ containing a specific vertex $s \in P$ that are connected, and whose vertex count $|P'| \leq k$ is bounded by the order k, is bounded from above by

$$\sum_{j=1}^{k-1} 2^{c(k-j)} = \sum_{j=1}^{k-1} \big(\underbrace{2^c}_{=:C}\big)^{k-j} = \sum_{j=1}^{k-1} C^{k-j} < \sum_{j=0}^{k-1} C^j \overset{C\geqq 2}{=} \frac{C^k - 1}{C-1} \leq C^k. \tag{11}$$

Proof. We will give a constructive proof by starting from a specific vertex and by adding further vertices to the current subgraph, moving along connected edges only.

Let a vertex $s \in P$ be given. We split the edges in equidistant levels with respect to s. To this end, let $K_s(j)$ be the set of terminal edges connecting any $v \in P$ to s via a shortest path of distance $d(s,v) = j$.

Consider now a possible subgraph G' with $s \in P'$. Let $K'_s(j)$ be the reduced set of edges of $K_s(j)$ for which only one of either associated vertices is in P', see Fig. 3 for a depiction of these sets. This set is the exploration boundary of G' at distance j with respect to s.

The cardinality of the power set of the reduced set of edges $K'_s(j)$ is $2^{|K'_s(j)|}$ for a level j. Therefore, we obtain $\sum_{j=1}^{k-1} 2^{|K'_s(j)|}$ possible sets by summing over all $k-1$ levels and ignoring that the number per level is actually not independent. With the upper bound on the vertex degree it follows that $|K'_s(j)|$ is bounded from above by $c|I_{j-1}|$ where I_{j-1} denotes the set of vertices added on level $j-1$. In Fig. 3 these are nodes designated with "1" and colored in dark gray. Furthermore, $|I_j|$ is bounded from above by $k-j$ because at least one vertex has to be added per level and there is already one root vertex. Putting it all together and using the partial sum of the geometric series results finally in (11). □

Hence, the sum of all possible subgraphs with at most k vertices only depends on the bond order k and the highest degree c over all vertices v in G. As we go over all vertices $s \in P$ as root vertices, the number of fragments scales as $O(M \cdot C^k)$.

4 Numerical Results

Now we present the results of our numerical experiments. This section is divided into three parts. In the first part, we look at the scaling behavior in terms of runtime

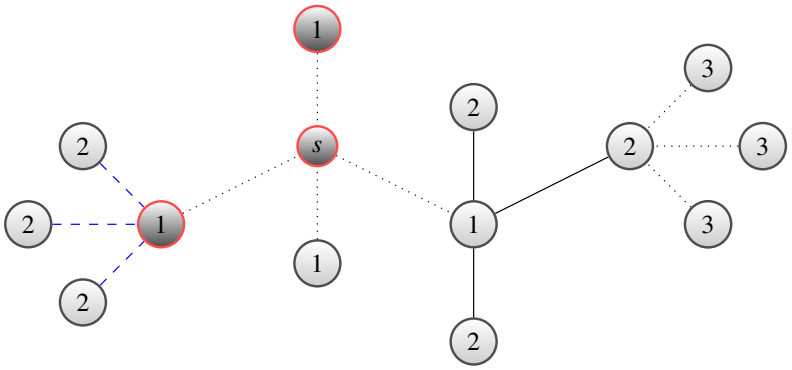


Fig. 3 Depiction of the reduced edge set $K'_s(2)$ with dashed lines for a given subgraph $G' \subset G$ consisting of vertices in dark gray with root vertex s. The vertices of the graph G are designated by the distance to s, all edges outside of the full edge $K_s(2)$ are dotted.

to assure linear complexity. In the second part, we give the approximate total energy for smaller molecules to indicate the good approximation quality of the approach. In the third and final part, we look at a large biomolecule and assess the applicability of the approach for large-scale calculations.

As approximate computational method for the electronic subproblems associated with the different fragments we have chosen the closed-shell Hartree Fock method with the Gaussian basis "6-311*G" set as implemented in MPQC [23]. We use evaluations of the total molecule as reference results (full HF) to compare the approximation error against.

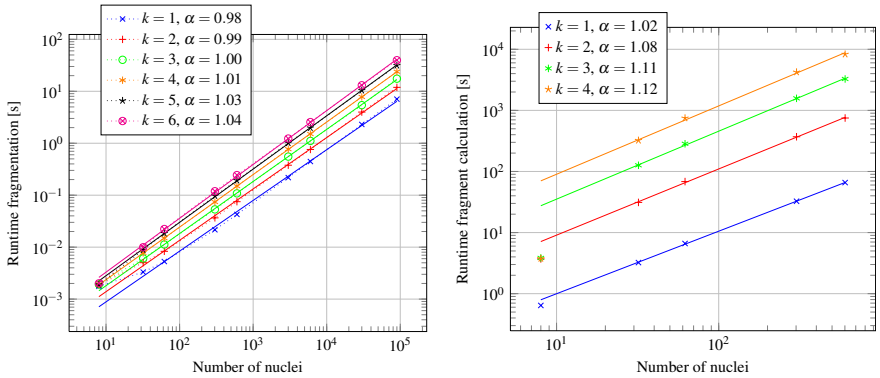
4.1 Scaling study

In the first part of this subsection we investigate the computational scaling behavior of our BOSSANOVA implementation with respect to the number of nuclei M and with respect to the truncation order k. From the theoretical considerations of the previous section, we here expect a linear scaling complexity with M.

To this end, we studied alkanes of varying length. In Fig. 4(a) the total runtime for the fragmentation procedure is given and in Fig. 4(b) the cumulative runtime for the calculation of all fragment problems is depicted. Both show linear scaling behavior with the number of nuclei M as expected. Additionally, we see that the time required for the calculation indeed increases polynomially with the truncation order.

Finally, we measure the crossover point for our ansatz—that is when the fragment calculations require less time than the reference calculation of the full system[8]. To this end, we use four solvers for the fragment problems in parallel and compare

[8] As can be seen from Fig. 4, the fragmentation times are negligible.



(a) Fragmentation (b) Calculation

Fig. 4 Runtimes for the fragmentation and subsequent calculation of the individual fragments for alkanes of increasing length and varying truncation order $k = 1, \ldots, 6$.

against the runtime of MPQC running on four processes for the reference calculation in Fig. 5. The respective crossover point is where the black curve intersects the other curves associated with the varying truncation order k.

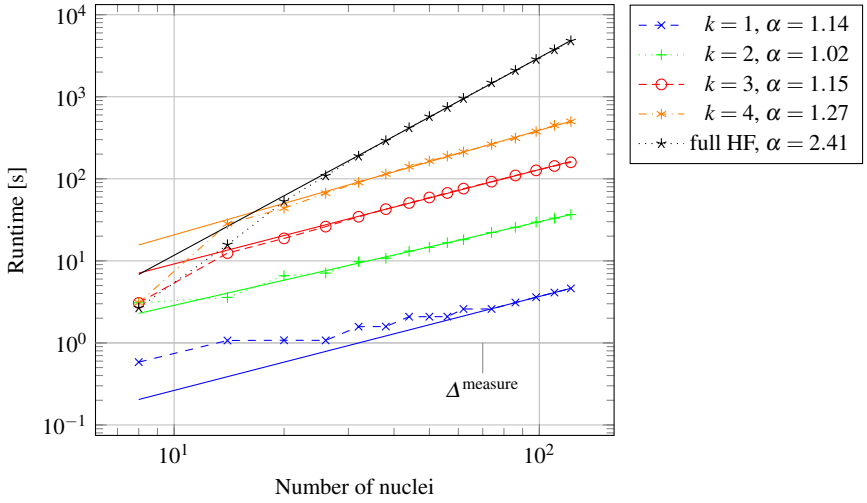


Fig. 5 Measured runtime of the calculation of alkanes of increasing length, via standard closed-shell HF and via BOSSANOVA for orders $k = 1, \ldots, 4$. Solid lines give linear regression fits to overall behavior.

We notice that at order $k = 4$ we obtain a crossover in runtimes at $M \approx 20$ which is roughly an order of magnitude in number of atoms, or three orders in total run

time, lower than that achieved by other linear-scaling schemes, e. g. ONETEP [34], see also [15].

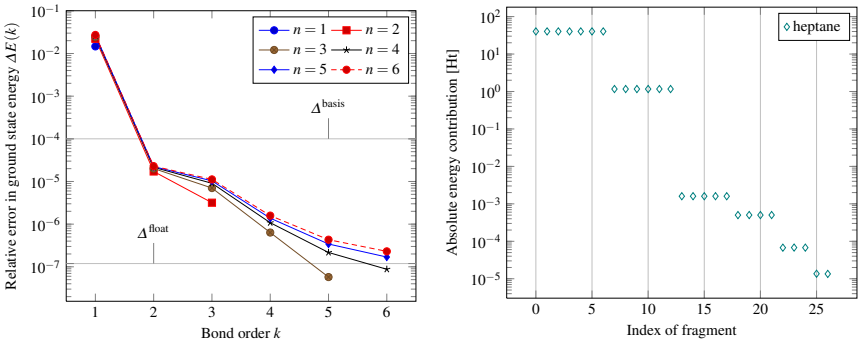
4.2 Qualitative study

In this section we investigate the approximation quality of the proposed approach. To this end, let us first give some remarks on what a threshold for a good approximation would be. HF calculations do not give the so-called correlation energy. However, due to the finite basis set they also never reach the true HF ground state energy but only an upper bound. For the employed "6-311G*" basis set we have estimated this finite basis set error (by employing even larger basis sets) to be 1.81×10^{-4} with respect to the true HF energy of alkanes. Hence, if we find the relative error $\Delta E(k)$ of the approximated energy E^{ANOVA} according to (10),

$$\Delta E(k) = \frac{E^{\mathrm{SCF}} - E^{\mathrm{ANOVA}}(k)}{E^{\mathrm{SCF}}}, \tag{12}$$

to be closer than $\Delta^{\mathrm{basis}} = 10^{-4}$ to the reference calculation E^{SCF}, we define the approximation to be good. As a second threshold value we use $\Delta^{\mathrm{float}} = 1.19 \times 10^{-7}$ as the output precision of values, i. e. below that value numerical rounding artifacts may appear.

In the following we give the numerical results for various chain molecules, namely alkanes, alkenes, alkines, and homologous chains consisting of boron and nitrogen. Let us remark that, for certain small lengths, we already reach the exact result for small truncation order k due to the nature of the telescopic sum.



(a) Alkanes (b) Heptane

Fig. 6 Approximation of the total ground state energy for various alkanes over the truncation oder k and absolute value of the energy contribution of each fragment of heptane sorted by increasing number of nuclei.

Table 1 Relative error $\Delta E(k,n)$ for increasing truncation order k and varying chain length n of the alkane molecule.

k	$\Delta E(n=3)$	$\Delta E(n=4)$	$\Delta E(n=5)$	$\Delta E(n=6)$
1	$2.47\cdot10^{-2}$	$2.60\cdot10^{-2}$	$2.67\cdot10^{-2}$	$2.72\cdot10^{-2}$
2	$2.02\cdot10^{-5}$	$2.16\cdot10^{-5}$	$2.24\cdot10^{-5}$	$2.29\cdot10^{-5}$
3	$7.01\cdot10^{-6}$	$9.06\cdot10^{-6}$	$1.03\cdot10^{-5}$	$1.12\cdot10^{-5}$
4	$5.95\cdot10^{-7}$	$1.08\cdot10^{-6}$	$1.35\cdot10^{-6}$	$1.55\cdot10^{-6}$
5	$8.50\cdot10^{-8}$	$1.91\cdot10^{-7}$	$3.06\cdot10^{-7}$	$4.26\cdot10^{-7}$
6	0.0	$6.38\cdot10^{-8}$	$1.53\cdot10^{-7}$	$2.13\cdot10^{-7}$

In Fig. 6(a) and Table 1 we give the relative error of the energy calculated for alkanes of length n with formula $C_{2n}H_{2n+4}$. We notice that we are below the estimated threshold Δ^{basis} already for $k=2$. Also, the error grows only very slowly for longer chains. Hence, the approximation works very well for these linear chain molecules, whose graph forms a tree and each edge represents only a single bond.

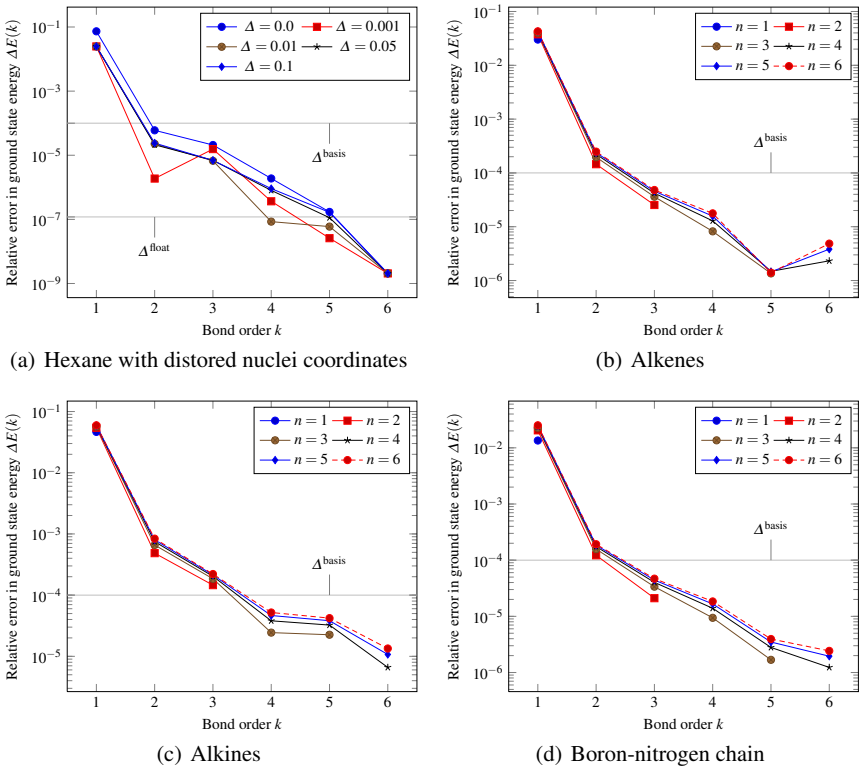
Furthermore, we depict the absolute value of the contribution to the total energy per fragment for heptane in Fig. 6(b). Due to the symmetry of the molecular system we clearly see levels of equal values in the graph. The difference between these levels closely follows the error obtained, e. g. 10^{-2} between level $k=1$ and $k=2$ and 10^{-3} between level $k=2$ and $k=3$. Hence, we feel that this can be taken as a rough error estimate when a full calculation is unavailable.

The approximation for hexane, alkenes and alkines, and boron-nitrogen chains of varying lengths with distored coordinates, higher bond degrees or different nuclei elements are depicted in Fig. 7.

We see that perturbation affects the approximation quality only negligibly. A stronger effect is seen with double and triple bonds as in alkenes and alkines or for different nuclei elements as with the boron nitrogen chain. However, we still reach the threshold Δ^{basis} at $k=3$ and notice that the decrease with chain length n is very small.

Moving on from these simple chain molecules to more complex bond graphs we come to molecular systems with aromatic rings. These are particularly difficult as the gain in energy due to the delocalized π-electrons is captured only when the complete ring is taken into account as a fragment. As an example we take naphthalene which consists of two interconnected aromatic rings and coronene which consists of six interconnected aromatic rings. In Fig. 8 we have calculated the approximative energy of these molecules two times: Once, we calculated the energy in the proposed fashion with increasing truncation order. The second time, however, we took the full cycles in the graph as extra fragments into account.

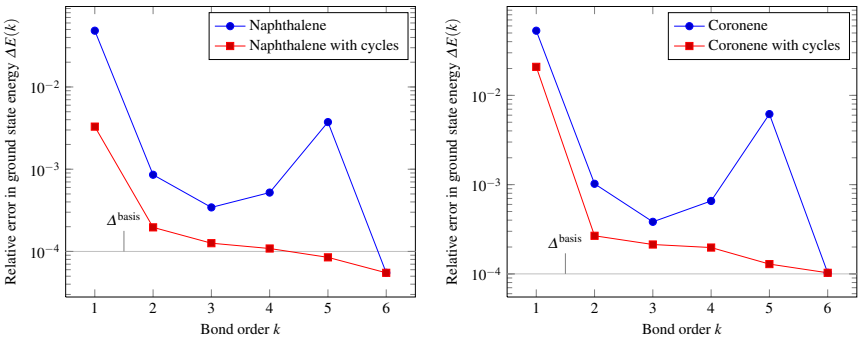
We immediately notice the effect: While with the first calculation the approximation error decreases up to $k=3$, it increases afterwards as higher-order fragments are strained due to the ring-like geometry of the full system. In the second calculation this decrease is absent although we never calculate the full system consisting of multiple interconnected rings. Moreover, we reach the threshold Δ^{basis} at around $k=5$.



(a) Hexane with distored nuclei coordinates (b) Alkenes

(c) Alkines (d) Boron-nitrogen chain

Fig. 7 Approximation of the total energy for hexane with nuclei coordinates under random perturbation of magnitude Δ, alkenes and alkines with double and triple bonds, and boron-nitride chain molecules of varying length n.



(a) Naphthalene (b) Coronene

Fig. 8 Relative error of the total energy for molecules with delocalized electrons over the truncation order k. In the second calculation cycles in the interaction graph are taken into account irrespectively of the truncation order.

4.3 Quantitative study

As an example of a truly large molecule we have chosen the interferon alpha (1ITF), taken from the Protein Data Bank [6] and amended it by hydrogens from topological knowledge via [22] that go undetected in the x-ray spectroscopy of the structure. The structure consists in total of 2698 nuclei.

Due to the larger number of nuclei a reference calculation is infeasible. Instead, we give in Fig. 9 the contributions to the telescopic sum from each individual fragment sorted by the number of nuclei. Each absolute energy value is given as a tiny dash in the figure that as a whole emphasizes certain levels of similar values, compare with Fig. 6(b). Also, we give exemplary fragments to each of the more prominent levels in Fig. 9(b).

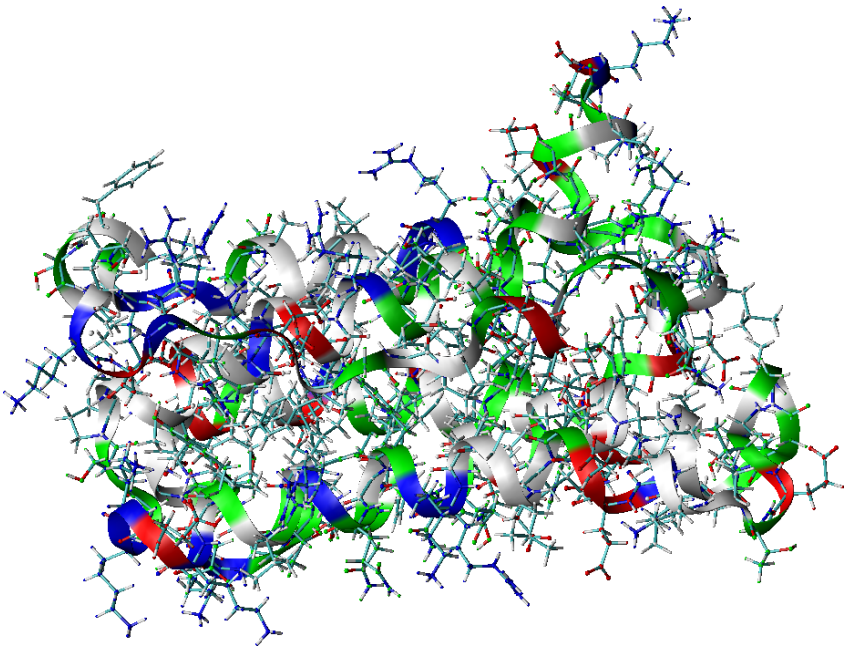
We notice a similar decay in the absolute magnitude as for alkane. This indicates a good approximation of the total energy of the structure. Judging from our previous remarks when investigating the approximation quality with alkane we see that the obtained ground state energy value of -68467.41 Ht is accurate to relative precision of 10^{-3} to 10^{-4}. This especially underlines the usefulness of the empirical potential approaches for these large biomolecules, see [32].

We remark that the cumulated solver runtime is 6.28 hours for this system. Hence, we see that our proposed scheme is especially well-suited to large molecules. Note furthermore that long-range Coulomb interactions can additionally be computed via one of the well-known schemes [4] in a first-order perturbation calculation [20] under the assumption that the wavefunctions do not change significantly anymore.

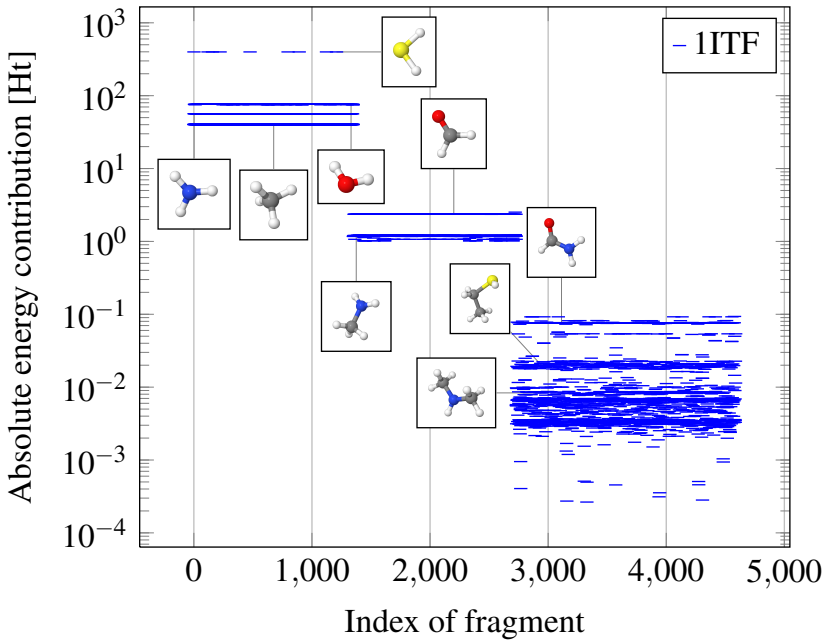
5 Concluding Remarks

In this article we presented the BOSSANOVA decomposition approach for the approximate solution to the electronic Schrödinger equation for a given molecular system. It involves an ANOVA series expansion of an electronic energy function in the framework of the Born-Oppenheimer molecular dynamics. A truncation of this series at a certain *bond order* and the elimination of certain further terms by a locality constraint of the electronic wavefunction plus some additional hydrogen saturation results in a set of fragments of the overall molecule. Now, each of the associated electronic subproblems may be solved with e.g. HF, CI, CC, or DFT methods. A proper combination of these solutions of the subproblems then leads to an approximate total ground state energy. This is an extension of the so-called additivity models which are well-known in chemistry.

We gave a description how this truncated BOSSANOVA expansion can be derived for any given graph. Furthermore we showed theoretically as well as practically that our new method indeed scales linearly with the number M of atoms in the overall problem. We gave numerical results for chain molecules where the obtained relative accuracy was well below 10^{-4} for $k = 3$, which is the relative precision



(a) Ball-stick-model



(b) Absolute contribution per fragment

Fig. 9 Ball and stick model of interferon alpha (PDB key: 1ITF) combined with a ribbon view of the main chain. The configuration is split up into fragments of up to $k = 3$ for which we give each's contribution to the total energy and examples of typical fragment subsystems.

of the reference calculation with respect to an infinite basis set. We also investigated aromatic systems with delocalized electrons where an inclusion of full cycles aids the approximation significantly to also achieve 10^{-4} relative precision. This is roughly the precision available to HF calculations with moderately sized basis sets.

Note that the impact of the neglected long-range Coulomb energy on the accuracy of the method and ways to recover this contribution is given elsewhere, see [20]. Note furthermore that our BOSSANOVA approach is not rid of empirical parameters due to the necessity to saturate dangling bonds with hydrogen in the fragmentation process. Since the typical bond lengths and angles of hydrogenated systems are well assessed by measurements, we hope that a careful collection of robust values into a database may enable a broad range of application for the BOSSANOVA method.

Let us also point out that our approach is trivial to parallelize since the evaluation of each fragment by an appropriate solver can be done independently, see [20]. Furthermore, since each fragment only contains a number of atoms which is roughly equal to the bond order k (neglecting hydrogen), the evaluation of the subproblems is possible already on very small machines with minimal memory requirements. Of course, also the memory cost scales only linearly. Thus, if the energy of a single fragment is calculated in seconds by e.g. a solver which is specifically tailored to the fast but precise evaluation of small and isolated systems, even a number of 10^5 or 10^6 fragments is within reach and the approximate total ground state energy evaluation of huge homogeneous molecular systems becomes computationally feasible. This has been shown by the calculation of the ground state energy of interferon alpha.

Finally, let us remark on how the BOSSANOVA method may be incorporated into a general coupling scheme of QM and MM. The BOSSANOVA fragmentation would be executed only in a given local domain, i.e. the active region where QM is locally needed. The resulting fragments are then forwarded to a suitable QM solver whereas the surrounding passive environment would not be fragmented but is directly passed on to a MM solver. Our BOSSANOVA scheme is closely related to conventional many-body potentials (however in an ab-initio fashion) with variable many-body order. Furthermore, due to the fragmentation process, the interface region is not sharply defined. Therefore, we believe that this approach also remedies the problems of energy and electron density leaking of other local coupling methods to a certain extent.

Acknowledgements This research was funded by the Deutsche Forschungsgemeinschaft (DFG) within the framework of the priority program SPP1324 and of the Collaborative Research Centre 1060 “The Mathematics of Emergent Effects” at the University of Bonn.

References

1. Abell, G.C.: Empirical chemical pseudopotential theory of molecular and metallic bonding. Physical Review B **31**(10), 6184–6196 (1985)

2. Amovilli, A., Cacelli, I., Campanile, S., Prampolini, G.: Calculation of the intermolecular energy of large molecules by a fragmentation scheme: Application to the 4-n-pentyl-4-cyanobiphenyl (5CB) dimer. Journal of Chemical Physics **117**, 3003–3012 (2002)
3. Antes, I., Thiel, W.: Adjusted connection atoms for combined quantum mechanical and molecular mechanical methods. Journal of Physical Chemistry A **103**(46), 9290–9295 (1999)
4. Arnold, A., Bolten, M., Dachsel, H., Fahrenberger, F., Gähler, F., Halver, R., Heber, F., Hofmann, M., Holm, C., Iseringhausen, J., Kabadshow, I., Lenz, O., Pippig, M., Potts, D., Sutmann, G.: Comparison of scalable fast methods for long-range interactions. Physical Review E **88**(6), 063,308 (2013)
5. Ayala, P.Y., Scuseria, G.E.: Linear scaling second-order Moeller-Plesset theory in the atomic orbital basis for large molecular systems. Journal of Chemical Physics **110**(8), 3660–3671 (1999)
6. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank (2000). URL http://www.pdb.org/
7. Brenner, D.W.: A second-generation reactive bond order (REBO) potential energy expression for hydrocarbons. Journal of Physics: Condensed Matter **14**, 783–802 (2002)
8. Challacombe, M.: A simplified density matrix minimization for linear scaling self-consistent field theory. Journal of Chemical Physics **110**, 2332–2342 (1999)
9. Collins, M.A., Deev, V.A.: Accuracy and efficiency of electronic energies from systematic molecular fragmentation. Journal of Chemical Physics **125**, 104,104 (2006)
10. Csyani, G., Albaret, T., Payne, M.C., De Vita, A.: Learn on the fly: A hybrid classical and quantum-mechanical molecular dynamics simulation. Physical Review Letters **93**(17), 175,503 (2004)
11. Daw, M.S., Baskes, M.I.: Embedded-atom method: Derivation and application to impurities, surfaces and other defects in metals. Physical Review B **29**(12), 6443–6453 (1984)
12. Deev, V., Collins, M.A.: Approximate ab initio energies by systematic molecular fragmentation. Journal of Chemical Physics **122**(15), 154,102 (2005)
13. Ercolessi, F., Adams, J.B.: Interatomic potentials from 1st-principles calculations - the force-matching method. Europhysics Letters **26**(8), 583–588 (1994)
14. Fonseca Guerra, C., Snijders, J.G., te Velde, G., Baerends, E.J.: Towards an order-N DFT method. Theoretical Chemistry Accounts **99**(6), 391–403 (1998)
15. Goedecker, S.: Linear scaling electronic structure methods. Reviews of Modern Physics **71**(4), 1085–1123 (1999)
16. Greengard, L., Rokhlin, V.: The fast multipole method for gridless particle simulation. Computer Physics Communications **48**, 117–125 (1988)
17. Gresh, N., Claverie, P., Pullman, A.: Theoretical studies of molecular conformation. Derivation of an additive procedure for the computation of intramolecular interaction energies. Comparison with ab-initio SCF computations. Theoretica Chimica Acta **66**, 1–20 (1984)
18. Griebel, M., Knapek, S., Zumbusch, G.: Numerical Simulation in Molecular Dynamics – Numerics, Algorithms, Parallelization, Applications. Springer-Verlag, Heidelberg (2007)
19. Hayes, M.Y., Li, B., Rabitz, H.: Estimation of molecular properties by high-dimensional model representation. Journal of Physical Chemistry **110**, 264–272 (2006)
20. Heber, F.: Ein systematischer, linear skalierender Fragmentansatz für das Elektronenstukturproblem. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn (2014)
21. Hoeffding, W.: A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics **19**(3), 293–325 (1948)
22. Humphrey, W., Dalke, A., Schulten, K.: VMD – Visual Molecular Dynamics. Journal of Molecular Graphics **14**, 33–38 (1996)
23. Janssen, C.L., Nielsen, I.B., Leininger, M.L., Valeev, E.F., Kenny, J.P., Seidl, E.T.: The Massively Parallel Quantum Chemistry Program (MPQC), Version 2.3.0. Sandia National Laboratories, Livermore, CA, USA (2008). URL http://www.mpqc.org/
24. Johnson III, R.D.: NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101 (2006). URL http://srdata.nist.gov/cccbdb

25. Kitaura, K., Ikeo, E., Asada, T., Nakano, T., Uebayasi, M.: Fragment molecular orbital method: an approximate computational method for large molecules. Chemical Physics Letters **313**, 701–706 (1999)
26. Laio, A., Van de Vondele, J., Rothlisberger, U.: A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations. Journal of Chemical Physics **116**(16), 6941–6947 (2002)
27. Li, X.P., Nunes, R.W., Vanderbilt, D.: Density-matrix electronic-structure method with linear system-size scaling. Physical Review B **47**, 10,891–10,894 (1993)
28. Liu, W.K., Karpov, E.G., Zhang, S., Park, H.S.: An introduction to computational nanomechanics and materials. Computer Methods in Applied Mechanics and Engineering **193**, 1529–1578 (2004)
29. Marx, D., Hutter, J.: Ab initio molecular dynamics: Theory and implementation. In: Modern Methods and Algorithms of Quantum Chemistry, *NIC Series*, vol. 1, pp. 301–440. Forschungszentrum Juelich, Deutschland (2000)
30. Maseras, F., Morokuma, K.: IMOMM - a new integrated ab-initio plus molecular mechanics geometry optimization scheme of equilibrium structures and transition-states. Journal of Computational Chemistry **16**(9), 1170–1179 (1995)
31. Parr, R.G., Yang, W.: Density-Functional Theory of Atoms and Molecules. Oxford Science Publications (1989)
32. Ponder, J.W., Case, D.A.: Force Fields for Protein Simulation. Advances in Protein Chemistry **66**, 27–85 (2003)
33. Sauer, J., Sierka, M.: Combining quantum mechanics and interatomic potential functions in ab initio studies of extended systems. Journal of Computational Chemistry **21**(16), 1470–1493 (2000)
34. Skylaris, C.K., Haynes, P.D., Mostofi, A.A., Payne, M.C.: Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. Journal of Chemical Physics **122**(8), 84,119 (2005)
35. Szabo, A., Ostlund, N.S.: Modern Quantum Theory - Introduction to Advanced Electronic Structure Theory. Dover Publications (1996)
36. Tersoff, J.: Modeling solid-state chemistry: Interatomic potentials for multicomponent systems. Physical Review B **39**, 5566–5568 (1989)
37. Van der Vaart, A., Gogonea, V., Dixon, S.L., Merz jr., K.M.: Linear scaling molecular orbital calculations of biological systems using the semiempirical divide and conquer method. Journal of Computational Chemistry **21**(16), 1494–1504 (2000)
38. Velde, G.T., Bickelhaupt, F.M., Baerends, E.J., Guerra, C.F., Van Gisbergen, S.J.A., Snijders, J.G., Ziegler, T.: Chemistry with ADF. Journal of Computational Chemistry **22**(9), 931–967 (2001)
39. Vreven, T., Morokuma, K.: On the application of the IMOMO (Integrated Molecular Orbital + Molecular Orbital) Method. Journal of Computational Chemistry **21**(16), 1419–1432 (2000)