BOYS LOCALIZATION AND PIPEK-MEZEY LOCALIZATION OF INTERNAL COORDINATES AND NEW INTERMOLECULAR COORDINATES IN TURBOMOLE

Malte von Arnim
TURBOMOLE Gmbh
Litzenhardtstrasse 19
76135 Karlsruhe
maltevonarnim@gmx.de

July 19, 2023

ABSTRACT

Local internal coordinates can be achieved with simplified Versions of the localization methods of Boys and of Pipek-Mezey, which are applied to delocalized coordinates. However, the new methods are, at the publication date of this article, only included in an **experimental version** of Turbomole. Excellent localization is achieved for linear or branched chains in molecules, but also for rings and even for multiply fused rings one obtains localization within a few Angström units.
The delocalized coordinates of TURBOMOLE consist of special, contracted redundant internal coordinates, which can be linear combinations of primitive torsions, of primitive bond angles or of primitive out of plane angles respectively. Bond stretches are taken individually. This contraction reduces the computational effort drastically, and presumably it also improves the convergence of the localization.
*The conzept of **localization** can be generalized to a set of n vectors in an Euklidian and real valued vector space of dimension $\mathbf{m}$, the basis vectors must be orthonormal, with $\mathbf{n} \geq \mathbf{2}$ and $\mathbf{m} \geq \mathbf{2}$ and $\mathbf{m} \geq \mathbf{n}$. The condition for this is that you can assign a location vector in an Euklidian and real valued $\mathbf{l}$-dimensional vector space with orthonormal basis vectors, $\mathbf{l} \geq \mathbf{1}$, to each basis vector of the n vectors mentioned above in a canonical way.*
Descriptively, localization means that for each vector of the set mentioned above the basis vectors with larger absolute values of the coefficients are only distributed over as small an area in the $\mathbf{l}$-dimensional space as possible. My simplified versions of the Boys algorithm and of the Pipek-Mezey algorithm have performed well for this task.
In the second part of the paper new internal coordinates for supermolecules consisting of several isolated subunits are presented. The degrees of freedom of the relative movements of the subunits are well described, and thus geometry optimizations of supermolecules converge much faster in some cases. The stability of the generation of the internal coordinates in TURBOMOLE is also improved considerably.
The new intermolecular coordinates are are part of the **official** version and can be switched on by user input.

Keywords translation coordinates · rotation coordinates · translation and rotation coordinates · coordinates for supermolecules · localization of vectors · localization of orthonormal vectors · localized vectors · Boys localized vectors · Pipek-Mezey localized vectors · local internal coordinates · geometry optimization · contracted internal coordinates · localized internal coordinates

1 Introduction

Local internal coordinates have well known advantages [1, 2, 3], but their construction requires a considerable programming effort [4]. Here I present an *entirely new* way of generating local internal coordinates: Delocalized coordinates [5] are relocalized using simplified versions of the method of Boys or of Pipek Mezey respectively. The additional code required essentially consists of the computation of the two-index integrals for the Boys algorithm and of calling the localization routines, which are contained in almost any quantum chemistry program package.
However, the computational effort grows with third power of the number of redundant internal coordinates, whereas our old method of obtaining local internal coordinates scales linearly [4]. This disadvantage is greatly mitigated by the contraction of the redundant internal coordinates in Turbomole: Instead of using the huge number of primitive, i.e. single, torsions, bond angles, and out of plane angles we use a much fewer number of contracted internal coordinates. All single torsions at one chemical bond are combined into a linear combination. Linear combinations are also taken for bond angles or out of plane angles. The coefficients are determined using calculations on fragments, which consist of a central atom and all its bond partners.

An almost perfect localization one obtains for linear or branched chains in molecules. Degrees of freedom of Torsions and bond distances are represented by a single torsion or by a single bond stretch. Bending modes are represented by linear combinations of bond angles or out of plane angles, with all angle coordinates involved having the same apex atom.
For rings or multiply fused rings (denoted as cages) we get localizations within a few Angström units, but these localized coordinates are generally mixtures of redundant internal coordinates of different type (bond stretches, bond angles,...), and of of different locations. For atoms bound directly to cages (at cage) one always obtains a single bond stretch, and about half of the bending modes are also well localized. The remaining bending modes for at cage atoms however show strong mixtures with modes of the cage.
To quality of the localization can be further improved by localizing mode separated delocalized coordinates. Mode separated delocalized coordinates contain as components only redundant internal coordinates from one out of four categories: bond stretches, internal coordinates for bending modes, torsions, or coordinates connecting the isolated sub-molecules of a supermolecule. These categories are expected to represent modes of different average strength, i.e. their average force constants drop sharply from group to group. However, separating the different categories requires additional computational effort and several hundred lines of code. A localization with very little computing time consumption is possible by using generalized natural coordinates. This type of internal coordinates combines the advantages of natural internal internal coordinates [1] and mode separated coordinates.
Good convergence of the localization is achieved for Boys and Pipek-Mezey after fifty iterations. An iteration involves a Jacobi-Rotation for each pair of delocalized coordinates.
For geometry optimizations and other applications, internal coordinates, i.e. the columns of Wilsons B-matrix [11], need not to be orthogonal. For this reason one can delete coefficients with smaller absolute values in the vectors of the localized internal coordinates, as long one doesn't come to close to linear dependency. With this procedure, the length of the "orthogonality tails" can be greatly reduced; average lengths of 3 to 1.5 Angström units are observed.
In order to be able to apply Boys's and Pipek-Mezey's methods to delocalized internal coordinates, these algorithms had to be greatly simplified. Because of this, the conzept of **localization** can be generalized to a set of n vectors in an Euklidian and real valued vector space of dimension $\mathbf{m}$, the basis vectors must be orthonormal, with $\mathbf{n} \geq \mathbf{2}$ and $\mathbf{m} \geq \mathbf{2}$ and $\mathbf{m} \geq \mathbf{n}$. The condition for this generalization is that you can assign a location vector in an Euklidian and real valued $\mathbf{l}$-dimensional vector space with orthonormal basis vectors, $\mathbf{l} \geq \mathbf{1}$, to each basis vector of the n vectors mentioned above in a canonical way.
Descriptively, localization means that for each vector of the set mentioned above the basis vectors with larger absolute values of the coefficients are only distributed over as small an area in the $\mathbf{l}$-dimensional space as possible. My simplified versions of the Boys algorithm and of the Pipek-Mezey have performed well for this task.
Both methods can also be formulated for complex valued vectors, with complex valued position vectors for the components in the case of the Boys method. [6].
In the second part of this article, new internal coordinates for the relative movements of the isolated submolecules in a supermolecule are described. For the new procedure, internal coordinates are first defined for the individual submolecules with the usual algorithm of Turbomole. Only then are internal coordinates generated for the relative movements of the submolecules. In this way, the standard algorithm in Turbomole is no longer disturbed by "unchemical" or bizarre" bonds between the submolecules. The generation of internal coordinates for becomes much more stable in this way, especially for dimers with large distances. Geometry optimizations with the new internal coordinates in some cases require far fewer cycles.

2 Localization of internal coordinates:

2.1 Theory:

Except for two important details, the next two sections are also included in our previous paper [4]. However, they are repeated here because they are essential for understanding the following sections.

2.1.1 Delocalized coordinates

The starting point of our method, the delocalized coordinates [5], is closely related to the redundant coordinates [7]. The delocalized coordinates are obtained by diagonalizing the positive semidefinite $\mathbf{BB^t}$ matrix, which is set up in a redundant set of local internal coordinates. This set comprises of all bond stretches, bond angles and torsional angles of the molecule, the primitive internals. The $\mathbf{B}$ matrix contains the derivatives of primitive internal coordinates q_i with respect to cartesian components (x,y or z coordinates of atoms) x_j,

$$B_{ij} = \frac{dq_i}{dx_j}. \tag{1}$$

All solutions C_j of the eigenvalue equation

$$(\mathbf{BB^t})\mathbf{C_j} = \lambda_\mathbf{j}\mathbf{C_j} \tag{2}$$

belonging to eigenvalues larger than zero are taken as "delocalized coordinates". All redundancies are thus eliminated since they are contained only in the discarded eigenvectors belonging to eigenvalues which are exactly zero. The delocalized coordinates are complete, i.e., they account for all molecular degrees of freedom, if the underlying set of local internal coordinates was chosen carefully.
In our method we reduce the size of the redundant set by contraction and selection of primitive internals which can be defined for a molecule; the details will be explained below (see section 2.1.2). This leads to a drastic reduction of the number of redundant coordinates and maintains the completeness.

2.1.2 Contration and selection of the redundant internal coordinates

This section describes our modification of BKD method [5]. The aim is to reduce the number of internal coordinates to keep the diagonalization of $\mathbf{BB^t}$ and the subsequent localization feasible even for large molecules.
The set of redundant internal coordinates obtained with the methods described below is also referred to as *basis coordinates* in the following text of this article. The reason is the close analogy to the basis functions in molecular orbitals.

Fragments, bond stretches, bond angles, and out-of-plane angles The central idea of our procedure is the definition of fragments of a molecule; fragment $\mathbf{i}$ consists of the "central" atom $\mathbf{i}$ and all its bond partners provided atom $\mathbf{i}$ is not a terminal atom. We then construct a complete set of internals for fragment $\mathbf{i}$, typically comprising of bond stretches and linear combinations of bond angles. The set of fragment internals is finally completed by appropriate (linear combinations of) torsions to get a redundant set of internals for the total molecule. In details our procedure works as follows:
For each fragment $\mathbf{i}$ we consider all bond stretches as defined by the topological matrix. Then compute the corresponding $\mathbf{BB^t}$ matrix and diagonalize it, all eigenvectors with an eigenvalue larger than a threshold (default 0.002) are kept. If the number of these vectors is smaller than $\mathbf{3N_i - 6}$, where $\mathbf{N_i}$ is the number of atoms in the fragment, add bending modes.
Take all bond angles with apex atom $\mathbf{i}$ and proceed as described for the stretches. If the local fragment is *planar*, [Fragments are considered as planar, if the product $\mathbf{i} \cdot (\mathbf{j} \times \mathbf{k})$ for normalized vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$ is less than a threshold (default 0.4) for every tripod in the fragment.] define further all out-of-plane angles with central atom $\mathbf{i}$. In the same manner as for the bond angles contracted out-of-plane angles are obtained, which are included in the redundant set.
To improve the estimation of force constants in geometry optimizations [1], it is necessary to take all primitive bond stretches of a fragment as redundant coordinates instead of the eigenvectors of the corresponding $\mathbf{BB^t}$.

Torsions We construct one contracted torsion coordinate for each bond in the molecule if at least one primitive torsion can be defined for that bond. All primitive torsions around a specific bond are weighted equally, i.e. their contraction coefficients are identical. Even for large molecules with a high average number of bond partners we get only a moderate number of torsions. For example for an Al_{147} cluster in C_1 symmetry we get only 1332 contracted torsions whereas the number of primitive torsions is 93820.

Collinear chains The prescription breaks down if the molecule contains a collinear chain, defined as a connected set of atoms with exactly to bond partners each and a bond angle close to 180° (threshold 155°). The chain is described by bond stretches and special internal coordinates for collinear bending [4]. To apply the above algorithm we remove the chain from the topology matrix and replace it by a new bond.

2.1.3 Mode separated delocalized coordinates

Many molecules of interest have modes of very different strength, the $(H_2O)_6$ cluster (see Figure 1), e.g.; has strong $O-H$ bonds and weak torsions around $O\cdots H$ hydrogen bonds. For such molecules it is of advantage to have a set of internal coordinates which do not mix such modes.
We achieve the separation of the different modes by a sequence of steps; the first step is to sort the redundant internal coordinates (see Sec. 2.1.2) into groups expected to have decreasing average force constants;

1. primitive bond stretches
2. bond angles, out-of-plane angles and special bends for collinear chains
3. torsions
4. all internal coordinates which involve weak bonds, e.g. hydrogen bonds

We want to obtain coordinates consisting only of members of a specific group $\mathbf{n}$ (as just defined) by diagonalizing $\mathbf{B_n}(\mathbf{B_n})^\mathbf{t}$. To keep the four types of coordinates mutually linear independent, we consider the redundant internals in the above order and project out the space which is spanned by all preceding groups,

$$|\acute{B}_i^n\rangle = |B_i^n\rangle - \sum_{m=1}^{n-1}\sum_{j=1}^{N_m} \bar{B}_j^m\rangle\langle\bar{B}_j^m|B_i^n\rangle; \tag{3}$$

B_i^n: columns of $\mathbf{B_n}$; $\acute{B}_i^n$: projected columns of $\mathbf{B_n}$; $\bar{B}_j^m$: jth column of the transformed B matrix of group m, Eq. 4 below; N_m: number of columns in the transformed B matrix of group m. We keep eigenvectors of $\acute{\mathbf{B}}_\mathbf{n}(\acute{\mathbf{B}}_\mathbf{n})^\mathbf{t}$ which belong to sufficiently large eigenvalues, the resulting coordinates are termed *mode separated delocalized coordinates*. This selection is the crucial step and cannot be done using a simple cutoff for the eigenvalues, this will be described in detail later.
To obtain the matrix $\bar{\mathbf{B}}_\mathbf{n}$ needed for the projection in Eq. 3 we transform the projected matrix $\acute{\mathbf{B}}_\mathbf{n}$ with the selected eigenvectors $\mathbf{C^n}$and eigenvalues $\mathbf{d}$ (see Sec. 2.1.3),

$$\bar{\mathbf{B}}^\mathbf{n} = \mathbf{d}^{-\mathbf{1/2}}(\mathbf{C^n})^\mathbf{t}\acute{\mathbf{B}}^\mathbf{n}. \tag{4}$$

If all eigenvectors of a given group are selected, we use the *individual coordinates of the group* instead of the the eigenvectors. Experience has shown that the separation of coordinate types does not work well if the atoms in a molecule are too closely connected, i.e., if there are centers with more than six bond partners or three membered rings are present. We apply the more robust simple delocalized coordinates (see Sec. 2.1.1) in these cases.

Selection of the eigenvalues Since the starting point for the mode separated coordinates is a redundant set of internal coordinates, many modes of a given molecule can be expressed in internal coordinates of different types. An example is the deformation mode of a Graphene sheet $C_{30}H_{15}$ perpendicular to its plane which can be a linear combination of out-of-plane angles or of torsions. We have to choose how many coordinates of each type we employ, i.e. how many eigenvectors are selected in each group. Of course the total number of Eigenvectors must be equal to the number of degrees of freedom. Obviously the best choice is the set of eigenvectors which yield the highest eigenvalues of $\mathbf{BB^t}$. To find the optimum set is difficult however because the selecting or discarding of a specific eigenvector in a certain group has a strong influence on all later groups due to the projection formula Eq. 3. If for example the program had selected an eigenvector consisting of out-of-plane angles for $C_{30}H_{15}$, a linear combination of torsions describing the same movement would be projected out, even if it yields a larger eigenvalue. To circumvent this dilemma we start our decoupling procedure using a high initial value of the selection threshold. If not enough eigenvectors can be found we reduce the threshold so that the largest rejected eigenvalue will be accepted. We repeat the decoupling procedure until enough eigenvectors are found or until the threshold gets to small. Our initial value is $2.0{*}eiglow$ and the lower boundary is $0.5 * eiglow$. $eiglow$ is the lowest eigenvalue of $\mathbf{BB^t}$ for the complete set of redundant internal coordinates.

2.1.4 Generalized natural coordinates

This type of internal coordinates is explained in detail in our previous paper [4]. The starting point for the generation of the generalized natural coordinates are the *natural internal coordinates*. These internal coordinates were introduced

by Pulay, Fogarasi, Pang and Boggs [1] and they were extended to many more cases by us [4]. We define the natural internal coordinates by a formula while in [1] they are given as a table for the most important cases.
Natural internal coordinates and are generally considered to be the internal coordinates best suited to geometry optimizations, transition state searches or other applications [1]. Natural internal coordinates are local and non redundant, i.e. exactly one natural coordinate is defined for each degree of freedom of the terminal atoms or rings mentioned below.
The simplest case for natural internal coordinates is a terminal atom for which a bond stretching mode and two modes for angular movements are defined. The first degree of freedom is represented by a single bond stretch coordinate, whereas the other two are accounted for in most cases by linear combinations of bond angles or out of plane angles, cf. sect. 2.1.2. In some cases, a torsion required. Natural internal coordinates for sets of several terminal atoms bonded to a common central atom are defined in an analogous manner. The other two important cases for natural internal coordinates are a ring connected through *one* atom with the rest of the molecule and a fused ring, which shares *a bond* with the rest of the molecule [1]. [1].
For clean accounting, terminal atoms, rings or fused rings are deleted from the molecule after the definition of the natural coordinates. By iterating this process, each molecule is described by natural internal coordinates as much as possible.
In more complicated cases, that is, when a molecule has multi-connected structures such as adamantane, a Graphene sheet, etc., an irreducible remainder remains for which no more natural internal coordinates can be defined.
The mode separated coordinates explained in sect. 2.1.3 are used as internal coordinates for this remainder of the molecule. The natural internal coordinates are placed as an additional first group in front of the other four groups in the formalism of the mode separated coordinates (see sect. 2.1.3).
The natural internal coordinates are always linearly independent of each other and of the internal coordinates of the remainder of the molecule. For this reason the natural internal coordinates are always used as individual internal coordinates by the formalism of the mode separated coordinates. So of course natural internal coordinates are not localized, since the are already perfectly local.

2.1.5 The simplified Boys localization

The key idea of the Boys localization [9] is to minimize the *self repulsion energy* functional $R(\phi_i)$ of a set of N_{MO} *molecular orbitals* ϕ_i.

$$R(\phi_i) = \sum_{i=1}^{N_{MO}} \int d\mathbf{r_1} \int d\mathbf{r_2}\, \phi_i(\mathbf{r_1})\phi_i(\mathbf{r_1})(\mathbf{r_1}-\mathbf{r_2})^2\phi_i(\mathbf{r_2})\phi_i(\mathbf{r_2}) \tag{5}$$

The orbitals ϕ_i are mappings from an $\mathbf{l}$ dimensional euklidian space, $\mathbf{l} \geq \mathbf{1}$, including the vectors $\mathbf{r_1}$, $\mathbf{r_2}$ or $\mathbf{r}$, $\mathbf{r} = (x_1, ..., x_i, ..., x_l)$, onto the real numbers (we don't consider complex valued functions here).The functional $R(\phi_i)$ becomes the smaller, the better the orbitals are localized. This means that the values of the functions ϕ_i that are significantly different from zero must lie in the smallest possible spatial region. The funtional $R(\phi_i)$ has to be finite, i.e. the absolute function values of the ϕ_i must decay sufficiently quickly towards infinity. The orbitals ϕ_i must also be orthonormal, i.e. for the scalar products the following equation must apply:

$$\int d\mathbf{r}\, \phi_i(\mathbf{r})\phi_j(\mathbf{r}) = \delta_{ij} \tag{6}$$

The expression $(\mathbf{r_1}-\mathbf{r_2})^2$ in the functional $R(\phi_i)$ is the square of the distance of the arguments in the l-dimensional space of the two orbitals. A short calculation shows, that minimizing $R(\phi_i)$ is equivalent to maximizing $B(\phi_i)$,

$$B(\phi_i) = \sum_{i=1}^{N_{MO}} [\int d\mathbf{r}\, \phi_i(\mathbf{r})\ \mathbf{r}\ \phi_i(\mathbf{r})\ ]^2 \tag{7}$$

Both functionals are invariant with respect to translation or rotation of the coordinate system. Minimization of $R(\phi_i)$ under the constraint of orthonormality is equivalent to finding an real valued unitary transformation $\mathbf{U}$, with $\mathbf{U}\mathbf{U}^\mathbf{t} = \mathbf{1}$, which minimizes $R(\phi_i)$ or maximizes $B(\phi_i)$. Without loss of generality $\mathbf{U}$ can be composed of two by two Jacobi rotations of the form:

$$\phi'_s \quad = cos\theta\ \phi_s + sin\theta\ \phi_t \tag{8}$$

$$\phi'_t \quad = -sin\theta\ \phi_s + cos\theta\ \phi_t$$

[1]There are also special natural internal coordinates for collinear chains, see sect. 2.1.2

Der optimum rotation angle θ is given by:

$$\begin{aligned} sin4\theta &= B_{st}/(A_{st}^2 + B_{st}^2)^{1/2} \\ cos4\theta &= -A_{st}/(A_{st}^2 + B_{st}^2)^{1/2} \\ & \quad 0 < \theta < \pi/2 \\ A_{st} &= \langle s|\mathbf{r}|t\rangle^2 - 1/4[\langle s|\mathbf{r}|s\rangle - \langle t|\mathbf{r}|t\rangle]^2 \\ B_{st} &= \langle s|\mathbf{r}|t\rangle \left[\langle s|\mathbf{r}|s\rangle - \langle t|\mathbf{r}|t\rangle\right] \end{aligned} \tag{9}$$

Any multiples of $\pi/2$ can be added to θ. The expression $\langle s|\mathbf{r}|t\rangle$ is an abbreviation for the integral $\int d\mathbf{r}\ \phi_s(\mathbf{r})\mathbf{r}\phi_j(\mathbf{r})$. A loop over all pairs of orbitals ϕ_i, ϕ_j, a *sweep*, is repeated until $B(\phi_i)$ stops increasing. It is not guaranteed however, that the global maximum is reached by this procedure.
From now on, a special form is assumed for the functions ϕ_i: They are numerical functions which only have function values at certain support points $\mathbf{r_i}$ (see Eq. 10) in an $\mathbf{l}$-dimensional space, $\mathbf{l} \geq \mathbf{1}$. The function values are the coeffizients of $\mathbf{n}$ real valued orthnormal vectors in an $\mathbf{m}$-dimensional Euklidian space, with $\mathbf{n} \geq \mathbf{2}$ and $\mathbf{m} \geq \mathbf{2}$ and $\mathbf{m} \geq \mathbf{n}$. The integrals in Eq. 7 now have the form:

$$\langle s|\mathbf{r}|t\rangle = \sum_{i=1}^{m} C_i^s |\mathbf{r_i}| C_i^t \tag{10}$$

The scalar products are

$$\int d\mathbf{r}\ \phi_s(\mathbf{r})\phi_t(\mathbf{r}) = \delta_{ij} = \langle s|t\rangle = \sum_{i=1}^{m} C_i^s C_i^t \tag{11}$$

In the further course this article it is only about vectors (r, s) which consist of the *expansion koefficients* of *internal coordinates* with regard to *redundant internal coordinates*. In the sections 2.1.1, 2.1.2, 2.1.3 and 2.1.4 these internal coordinates are described in detail.
The support point $\mathbf{r_i}$ for a specific redundant internal coordinate is their center of mass:

$$\mathbf{r_i} = \frac{1}{\sum_{j=1}^{N_{prim}} c_{i,j}^2} \sum_{j=1}^{N_{prim}} c_{i,j}^2 \sum_{k=1}^{N_{atoms}} \mathbf{a_{j,k}}/N_{atoms} \tag{12}$$

N_{prim} is the number of primitive internal coordinates in a given redundant internal coordinate (see Sec. 2.1.2). N_{atoms} is the number of atoms in the primitive internal coordinates, which must all have the same type, their expansion coefficients are the $c_{i,j}$. $\mathbf{a_{j,k}}$ is the location vector of a certain atom contributing to a primitive internal coordinate.
In this paper, as in almost all other applications, the Boys localization is done in the usual cartesian space, $\mathbf{r}$ =X,Y,Z, i.e. $\mathbf{l} = 3$ (see above). For this reason, the scaling with increasing problem size is usualy given as $\mathbf{n^3}$. But of course the computation time also increases linearly with $\mathbf{l}$.
Two further methods are described below for defining the support points for the numerical functions ϕ_i.

Predefined points for bond stretches and torsions For bond stretches, the support points are on the bond axis, and for torsions they are on the bond around which the ligands rotate. The points of both types of coordinates are a distance 0.1 times bond length from the center of the respective bond. The support points of the bond stretches are always shifted from the center of the bond in the direction of the atom with the larger index, and the support points of torsions are shifted in the opposite direction. This should achieve a good separation of bond stretches and torsions, whose reference points are often very close together if they are calculated according to equation 12.
All support points for bond angles and out of plane angles are placed at the location of the apex atom of the respective redundant internal coordinate.

Points for bond angles and out of plane angles on a sphere The reference points of bond stretches and of torsions are defined exactly as described in the last paragraph. However, all support points of bond angles and of out of plane angles at a certain center see sec. 2.1.2 lie on a spere with the radius R_{min} around the location of the common apex atom. R_{min} is defined as $0.1 \times \sqrt{3}$ times the length of the shortest bond at the center.
The corners, face centers and edge centers of a cube are now projected onto the sphere. The points thus obtained on the sphere are then occupied with the reference points of the bond angles and out of plane angles. First all corners, then all face centers and then all edge centers are populated one after the other, until a place is found for the support point of every angle.

2.1.6 The simplified Pipek-Mezey localization

The orthonormal functions from section 2.1.5 can also be imagined as a linear combination of N_{bf} suitable basis functions $\mu(\mathbf{r})$. The exact nature of this basis functions is not important here.

$$\phi_i(\mathbf{r}) = \sum_{\mu=1}^{N_{bf}} c_{i,\mu}\mu(\mathbf{r}) \tag{13}$$

Each basis function $\mu(\mathbf{r})$ is assigned to a specific atom A. The physical nature of the N_A atoms is not important here. You can think of them simply as labels, which are used to classify the basis functions into groups placed in a specific place. The overlapp matrix $\mathbf{S}$ contains the overlap intgrals over the basis functions, these are generally not orthogonal to one another:

$$S_{\mu,\nu} = \int d\mathbf{r}\ \mu(\mathbf{r})\nu(\mathbf{r}) \tag{14}$$

The most important quantity of the Pipek-Mezey localization method, the *Mulliken atomic population* of an Orbital $\phi_i(\mathbf{r})$ for an atom A, is calculated from the overlap matrix and from the expansion coefficients of equation 13:

$$Q_A^i = \sum_{\nu\epsilon A}\sum_{\mu}^{N_{bf}} 2c_{i,\mu}c_{i,\nu}S_{\mu,\nu} \tag{15}$$

The first sum runs over all basis functions assigned to atom A. The underlying idea of the localization according to Pipek and Mezey [10] is that the populations of the N_{MO} individual orbitals $\phi_i(\mathbf{r})$ should be distributed over as few atoms as possible. The matching localization criterion for this is that

$$R(\phi_i) = N_{bf}^{-1}\sum_{i=1}^{N_{MO}}\sum_{A=1}^{N_A}(Q_A^i)^2 \tag{16}$$

is at its maximum. This criterion does not explicitly consider the positions of the atoms. However, informations about the spatial distribution of the atoms are contained in the overlap matrix.
The *atomic population operator* P_A is important for the further steps of the formalism. Applied to an orbital ϕ_i it generates the population of atom A.

$$P_A = \sum_{\mu\epsilon A}(1/2)\{|\tilde{\mu}\rangle\langle\mu| + |\mu\rangle\langle\tilde{\mu}|\} \tag{17}$$

$$|\tilde{\mu}\rangle = \sum_{\nu}(\mathbf{S^{-1}})_{\nu,\mu}|\mu\rangle$$

$$Q_A^i = \langle i|P_A|i\rangle$$

With the above definitions, the localization criterion 16 becomes:

$$P(\phi_i) = N_{bf}^{-1}\sum_{i=1}^{N_{MO}}\sum_{A=1}^{N_A}[\langle i|P_A|i\rangle]^2 \tag{18}$$

A strong resemblance can be seen with the corresponding criterion for the Boys localization, see Eq. 7. Only the components of the position vector $\mathbf{r}$ have been replaced by the projection operators for the individual atoms A.
The determination of the optimal rotation angle between two orbitals, ϕ_s and ϕ_t works the same as with the Boys localization. Only the terms A_{st} and B_{st} have to be replaced by the following expressions:

$$A_{st} \quad = \sum_{A=1}^{N_A}\langle s|P_A|t\rangle^2 - 1/4[\langle s|P_A|s\rangle - \langle t|P_A|t\rangle]^2 \tag{19}$$

$$B_{st} \qquad = \sum_{A=1}^{N_A}\langle s|P_A|t\rangle\,[\langle s|P_A|s\rangle - \langle t|P_A|t\rangle]$$

Again, the strong resemblance to the corresponding expressions for the boys localization can be seen, see. Eq. 9. The computation time increases with the third power of the problem size, and there is no further dependency on the dimension of the domain of the orbitals ϕ_i as for the boys localization, see sec. 2.1.5.
My simplification of Pipek-Mezeys Method now consists in replacing the overlap matrix, see Eq. 14, with an identity matrix. The matrix elements of the projection operator P_A become so:

$$\langle i|P_A|i\rangle = \sum_{\nu\epsilon A} c_\nu^i c_\nu^i \quad = Q_A^i \tag{20}$$

$$\langle s|P_A|t\rangle = \sum_{\nu\epsilon A} c_\nu^s c_\nu^t$$

One recognizes that the basis funtions $|\mu\rangle$ are no longer included in the formalism. For this reason this simplified localization scheme would be applicable to any set of orthonormal vectors ϕ_i (see below). The algorithm scales as $\mathbf{n^3}$ with the system size for all applications.
The sum in eq. 20 runs over all coefficients assigned to a specific atom A. To divide the coefficients into groups associated with each "atom", is the only remaining freedom of choice for the user.
For the *redundant internal coordinates* (see sect. 2.1.2) I have chosen the following subdivision scheme: Each bond length, each torsion and each coordinate for collinear bending is assigned to a separate atom. The other atoms contain groups of bond angles or out of plane angles, the members of a certain group must have the same apex atom.
By replacing the overlap matrix with an unit matrix (see above), *all information about the basis functions and the spatial distribution of the atoms has been lost*. In my applications, however, I have found that the quality of localization for the Boys method or for the Pipek-Mezey method is comparably good (see sect. 2.3.3).

2.1.7 Further aspects of the localization

Truncation of the orthogonality tails For geometry optimizations or for freezing individual degrees of freedom, the coefficient vectors of the localized coordinates or the columns of Wilsons $\mathbf{B}$-matrix for the localized coordinates (see sect. 2.1.1) do not have to be orthogonal to each other [1]. Therefore, one can truncate the so called *orthogonality tails* as long as the set of localized coordinates does not become linearly dependent as a result. The orthogonality tails result from the constraint of the localization (see Eq. 6) that the coefficient vectors of the localized coordinates must be orthonormal. In my implementation all coefficients whose absolute values are smaller than a given threshold are deleted.
A rough measure of the closeness to the linear dependency is the smallest Eigenvalue of the $\mathbf{BB^t}$-matrix (Marco Häser, private communication), where $\mathbf{B}$ is Wilsons B-matrix (see sec. 2.1.1) for the truncated localized coordinates. This value should not drop about a factor two when the tails are cut off.
After deleting the negligible coefficients the vectors of the localized coordinates are renormalized.
The localization itself does not change the eigenvalue spectrum of $\mathbf{BB^t}$ because it is a unitary transform.
The B-matrix elements for linear combinations of redundant internal coordinates (see sect. 2.1.1, 2.1.2, 2.1.3, 2.1.4) are calculated as:

$$B_{ij} = \sum_{k=1}^{N_{red}} C_k B_{ij}^k. \tag{21}$$

N_{red} is the number of redundant internal coordinates, B_{ij}^k is the B-matrix element of a specific redundant internal coordinate (see eq. 1 and sect. 2.1.2) and C_k is the associated coefficient. The coefficient vectors for natural internal coordinates (see sect. 2.1.4) and for redundant internal coordinates from linear independent groups (see sect. 2.1.3) are unit vectors.

Theoretical limit for the localization The maximum value that the squared coefficient can have for a given redundant internal coordinate with index j in a localized internal coordinate with index i is given by:

$$\max\left(C_j^i\right)^2 = 1 - \sum_{l=1}^{N_{zero}} \left(C_j^l\right)^2 \tag{22}$$

This equation follows from the orthonormality condition for the eigenvectors of the respective $\mathbf{BB^t}$-matrix (see sect. 2.1.1, 2.1.2, 2.1.3 and 2.1.4). N_{zero} counts the eigenvalues of $\mathbf{BB^t}$ which are exactly zero. The eigenvectors belonging to these eigenvalues are discarded to get rid of redundancies (see sec. 2.1.1, 2.1.3) and are therefore not included in the localization. Due to the renormalization of the truncated localized coordinates the theoretical limit no longer applies strictly if the cut off threshold is greater than 0 (see above).

Details of the implementation So far localization is only implemented for C_1-Symmetry.
Localization is turned off for molecules with coordination numbers greater than six and also for molecules containing three-memberered rings. Such constellations typically occur for metal clusters for which the classical localization of the molecular orbitals does not work well either.
The four different groups of mode separated coordinates (see sec. 2.1.3) are localized separately. Groups for which all eigenvectors have been selected are not localized, because these groups already consist of perfectly local internal coordinates, e.g. bond stretches, contracted torsions with a common axis or contracted bond angles with a common apex atom, see section 2.1.2.
Generalized natural coordinates are identical to mode separated coordinates, except that they contain the natural internal coordinates as an additional first group (see sect. 2.1.4). These are also perfectly local and therefore not localized.

When localizing internal coordinates for supermolecules generated with the method described in section 3, the procedure is as follows: The internal coordinates for the individual sub-molecules are localized separately using the methods described so far. The internal coordinates for the relative movement of the sub-molecules among each other are not localized.

2.2 Technical Details:

My localization of internal coordinates has been implemented in the Turbomole release of 11 November 2021, the version is still **experimental** and not yet part of the official Turbomole package. The routines boys.f and pipekmezey.f from the Turbomole package were used for the localization. All Turbomole source files were compiled with the Intel Fortran Compiler.
My example molecules (see figures 1 to 8) were created with the input generator *Tmolex* from Turbomole. I have taken important building blocks for the example molecules from *ChemSpider* or from the Turbomole test suite. All structures have been optimized with the Turbomole default: This is the RIDFT method with def-SV(P) basis sets for each atom and with the BP86 density functional (see Turbomole manual Version 7.4 and references therein).
In the examples $(H_2O)_6$-cluster, $(H_2O)_6$-cluster-chain and Fructosyl-nistose-$(H_2O)_3$, all intermolecular hydrogen bonds and all intramolecular hydrogen bonds were defined before the internal coordinates were generated. The criteria 2,3,5 and 6 of section 3.1 were used to detect hydrogen bonds. However, all hydrogen bonds were labeled as weak bonds. Therefore all basis coordinates containing an hydrogen bond were sorted into the group of "weak" coordinates when generating mode separated coordinates (see section 2.1.3).
All informations about the test molecules is available via my email address (see front page).
I have used a maximum of 50 sweeps (see sect. 2.1.5) for all test calculations to examine CPU-times and the quality of the localization (see tables 3, 4,5 and 6) and also for all test calculations for the truncation of the localized coordinates (see tables 7, 8, 9 and 10).
coordinates. For the test calculations on the localization quality, I used the value of 0.1 as the cut off criterion for the tails of the localized coordinates.

2.3 Results and Discussion:

2.3.1 Convergence of the localization for the different choices of the reference points

As example molecules, I have chosen the $(H_2O)_6$-cluster for table 1 and $C_{30}H_{13}$-prop-phen-naph for table 2, see figures 1 and 7. The tables show the number of sweeps required until the localization converges with the respective method. A sweep means performing one Jacobian rotation (see sec. 2.1.5) for each pair of vectors to be located in succession. For each specified number of sweeps in the header of tables 1 and 2, the maximum rotation angle (in Radians) in the Jacobi rotations of the last performed sweep is specified in each section of the tables. Localization is considered converged when the maximum rotation angle is less than 10^{-9} Radians. Therefore, on convergence, the actual number of sweeps performed is less than the number at the the top of the relevant column.
In the further rows in each section of tables 1 and 2 are given the average population radius and the average standard deviation of the distances of the populations $P_{\alpha,i}$ from their centroid $\vec{Z}_\alpha$ (see below) for each localized coordinate α. Both quantities are given in Å. The population radius R_α is calculated as:

$$\begin{aligned} R_\alpha &= \sum_{i=1}^{Natoms} P_{\alpha,i} |\vec{Z}_\alpha - \vec{r}_i| \\ \vec{Z}_\alpha &= \sum_{i=1}^{Natoms} P_{\alpha,i} \vec{r}_i \\ P_{\alpha,i} &= (B_{\alpha,ix}^{**2} + B_{\alpha,iy}^{**2} + B_{\alpha,iz}^{**2})^{1/2}/N \\ N &= \sum_{j=1}^{Natoms} (B_{\alpha,jx}^{**2} + B_{\alpha,jy}^{**2} + B_{\alpha,jz}^{**2})^{1/2} \end{aligned} \tag{23}$$

$P_{\alpha,i}$ is the normalized population computed from the three elements of Wilsons B-matrix $(B_{\alpha,ix}, B_{\alpha,iy}, B_{\alpha,iz})$ at an atom with index i in a localized coordinate with index α. $\vec{Z}_\alpha$ is the centroid of a localized coordinate α computed from the atomic populations $P_{\alpha,i}$ and from the atomic position vectors $\vec{r}_i$.
Of the two localization variants tested, the Pipek-Mezey method shows the best overall convergence. For the boys localization, three methods were tested to define the reference points of the redundant internal coordinates, see sec. 2.1.5. The first variant *centers of mass* converges a bit slower than the third variant. The second variant, on the other

hand, shows the slowest convergence for both examples. I still chose *centers of mass* because this method seems to be the most natural and also preserves the symmetry of the molecule. This allows localization to be implemented with symmetry, which is impossible for the other two variants.
The first variant is therefore used for all further Boys localizations carried out for this publication. For the Boys localization, the parameters mean population radius and mean sigma population are almost identical for all three variants tested. The localization quality is therefore equally good for all three variants (cf. sect. 2.3.3), as far as one can conclude from the two parameters.
From the data in both tables it can be clearly seen that the distribution of the populations over the atoms from fifty performed sweeps onward practically no longer changes. This applies even though the maximum rotation angle in the fiftieth sweep is still up to 3.3 % of the upper limit of $\pi/4$. Nevertheless, the localizations for all other tables were carried out with a limit of 50 cycles. I think this is justified because because the characteristic quantities for the atomic populations of the resulting internal coordinates do not change after 50 cycles.

Table 1: Convergence of the Localization for the $(H_2O)_6$-cluster

Max. No. of Sweeps	10	20	50	100
Boys, points are the centers of mass				
max. rotation angle	0.10	0.71-3	<1.00-9	-
mean population rad.	1.25	1.25	1.25	-
mean sigma population	0.77	0.77	0.77	-
Boys, predefined points for bond stretches and torsions				
max. rotation angle	0.90-1	0.28-1	0.40-7	<1.00-9
mean population rad.	1.27	1.26	1.26	1.26
mean sigma population	0.78	0.78	0.78	0.78
Boys, points for bond angles and out of plane angles on a sphere				
max. rotation angle	0.75-1	0.49-3	0.18-8	<1.00-9
mean population rad.	1.26	1.26	1.26	1.26
mean sigma population	0.78	0.78	0.78	0.78
Pipek-Mezey				
max. rotation angle	0.71-2	0.96-4	<1.00-9	-
mean population rad.	1.31	1.31	1.31	-
mean sigma population	0.82	0.82	0.82	-

2.3.2 Discussion of CPU-times

The measured CPU-times for the two example molecules $C_{30}H_{15}$-chain-corannulene ($C_{30}H_{15}$) and Fructosyl-nistose-$(H_2O)_3$ (Fru) are given in table 3. The CPU-times for the generation of the internal coordinates and for the subsequent localization are given. The abbreviations after the molecule designations indicate the type of localized internal coordinates. Delocalized coordinates (see sect. 2.1.1) are called **DL**, mode separated coordinates are denoted by **MS** (see sect. 2.1.3) and generalized natural coordinates are represented by **GN** (see sect. 2.1.4).
Both the localization and the step mentioned below in generating the internal coordinates scale with the cube of the system size.
The localization of the delocalized coordinates requires a multiple of the computing time for the generation of the internal coordinates for both the Boys-method and the Pipek-Mezey method. This could become a problem in practical applications. The CPU-times for the localization of the MS coordinates are of the same order of magnitude as the CPU-times for the internal coordinates, whereas the computing times for localizing the GN coordinates are very short, especially for the Boys-method. The explanation for this behavior is: With delocalized coordinates a large set of vectors must be localized together, whereas with the mode separarated coordinates the large set of vectors is divided into three roughly equal parts (see sect. 2.1.3), which are localized separately. In the case of the generalized natural coordinates, the dimension of this three subsets is significantly smaller than for the mode separated coordinates. This is because natural internal coordinates could be defined for a considerable part of the two test molecules, and natural internal coordinates are not localized (see section 2.1.4).
The CPU-times for the Pipek-Mezey method are consistently greater than those for the Boys-Method, because the update of the the orbital transformation matrix in each Jacobi rotation is more complex [10].

Table 2: Convergence of the Localization for $C_{30}H_{13}$-prop-phen-naph

Max. No. of Sweeps	10	20	50	100	200	500	1000	2000
Boys, points are the centers of mass								
max. rotation angle	0.12	0.74-1	0.26-1	0.60-2	0.10-3	0.63-5	0.62-7	<1.0-9
mean population rad.	1.03	1.02	1.01	1.01	1.01	1.01	1.01	1.01
mean sigma population	0.61	0.59	0.57	0.57	0.57	0.57	0.57	0.57
Boys, predefined points for bond stretches and torsions								
max. rotation angle	0.10	0.33-1	0.18-1	0.40-2	0.13-2	0.31-4	0.40-7	<1.0-9
mean population rad.	1.02	1.01	1.01	1.01	1.01	1.01	1.01	1.01
mean sigma population	0.59	0.58	0.58	0.58	0.58	0.58	0.58	0.58
Boys, points for bond angles and out of plane angles on a sphere								
max. rotation angle	0.86-1	0.30.1	0.11-1	0.77-2	0.77-3	0.83-6	<1.0-9	-
mean population rad.	1.02	1.01	1.01	1.01	1.01	1.01	1.01	-
mean sigma population	0.61	0.59	0.58	0.58	0.58	0.58	0.58	-
Pipek-Mezey								
max. rotation angle	0.58-1	0.11-1	0.28-2	0.82-2	0.35-4	<1.0-9	-	-
mean population rad.	1.17	1.18	1.18	1.17	1.17	1.17	-	-
mean sigma population	0.70	0.70	0.70	0.69	0.69	0.69	-	-

The computation time for the generation of the mode separated coordinates is significantly greater than the computing time for the delocalized coordinates. This applies, because in the second case the creation of the mode separated coordinates (2.1.3) is added to the diagonalization of the $\mathbf{BB^t}$-matrix of the redundant internal coordinates. [2] In the case of generalized natural coordinates, the computing times are a little shorter again, because in this case mode separated coordinates only have to be generated for the irreducible residues of the molecules (see sect. 2.1.4).
The measured Wall clock times deviate only slightly from the CPU-times and are therefore not discussed further here.

Table 3: CPU-times for the localization in seconds

Molecule/Method	$C_{30}H_{15}$ **DL**	$C_{30}H_{15}$ **MS**	$C_{30}H_{15}$ **GN**	Fru **DL**	Fru **MS**	Fru **GN**
Boys	2.37	0.26	0.04	4.76	0.33	<0.01
Pipek-Mezey	6.70	0.98	0.13	7.04	0.38	0.02
Internal coords.	0.12	0.22	0.28	0.12	0.18	0.25

2.3.3 The quality of the localization

In this section, the properties of the localized coordinates will be discussed in more detail. I have examined the localization using three example molecules. The first table (4) is for $(H_2O)_6$-cluster-chain (see fig. 2). The second table (5) is for Adamantane-chain (see fig. 4) and the third table (6) is for the molecule 3-spiro-rings-chain (see fig. 3). The data in the upper halves of the tables were generated with the Boys-localization (see sect. 2.1.5), and the data in the lower halves of the tables were obtained with the Pipek-Mezey-localization (see sect. 2.1.6).
There are basically three classes of localized coordinates. The first class is referred to as **in cage**, which means that the coordinate in question is mainly located in a cage. In addition to the structures commonly referred to as cages, e.g. Adamantane or Graphene sheets, rings are also counted among the cages here.
The second class of localized coordinates I call **at cage**. This means the degrees of freedom of atoms bound directly to a cage, so the bond stretch, which connects the *at cage* atom to the cage, and two angle coordinates. These are typically bond angles or out of plane angles (see sect. 2.1.2), which always have their apex atom in the cage.
The last class **in chain** contains all degrees of freedom of chains, these can be linear or branched chains. The chains

[2] This diagonalization is only required to get a starting value (*eiglow*) for the iterative generation of the mode separated coordinates.

can be attached to a cage or they can be a bridge between two or more cages. An exact definition for "chain" is that it must not contain any ring-like structures.
Each of the above three classes of localized coordinates can be further diveded into subgroups based on the type of those coordinates. The types considered in the tables 4, 5 and 6 are: bond stretch *stre*, bond angle *bend*, out of plane angle *outp*, and torsion *tors* (see sect. 2.1.2). The classification of a specific localized coordinate in its subgroup is based on its coefficient vector, the basis coordinate that has the coefficient with the largest absolute value is decisive. The appropriate subgroup is determined on the basis of the spatial position in the molecule and the type of this basis coordinate. For bond angles and out of plane angles the location of the apex atom counts. If the apex atom is connected directly to a cage or if it is further away from a cage, these types of coordinates are *in chain*. In the case of torsions, the location of the bond axis determines the classification into one of the three classes. If this bond axis is connected directly to cage, the torsions counts as *in chain*. Torsions whose bond axis is in a cage usually count as *in cage*. But if at least 50 % of the population (see eq. 23) of such a torsion belongs to *at cage* atoms, that torsion counts as *at cage*.
The following characteristic quantities are given in the tables 4, 5 and 6 for each of the subgroups mentioned above, these quantities are averaged over all members of a specific subgroup:

No. The number of localized coordinates in a given subgroup.

% Type If you sum up in the coefficient vector of a given localized coordinate the coefficient squares of all basis coordinates that have the same type (see above) as this localized coordinate, the quantity % Type results.

Tail length This means the maximum distance in Å units that a population at a specific atom with a magnitude greater than 0.001 has from the center of mass of a given localized coordinate. The size of the populations and their distance from the center of mass are calculated using the formulas given in section 2.3.1.

% Local A so called local group is defined for each localized coordinate. The structure of this local group depends on the type (see above) of a given localized coordinate. For bond stretches, the two atoms of the bond are the local group. Local groups for bond angles and out of plane angles consist of the apex atom and all atoms bound to the apex atom. For torsions one takes the atoms of the bond axis and their ligand atoms.
In general, a specific localized coordinate does not have all of its population (see sect. 2.3.1) in its local group. % Local now reports the percentage of the population that is in its local group for this localized coordinate.

If a localized coordinate has 100 % of its population in its local group and also has 100 % of a certain type, it is said to be *perfectly localized*.
Each of the three tables 4, 5 and 6 contains the data for delocalized coordinates (see sect. 2.1.1) in its left part and the data for mode separated coordinates (see sect. 2.1.3) in its right part. All columns with data for delocalized coordinates are marked with a **1** and all columns for mode separated coordinates are marked with a **2**.

Global tendencies Because the tables contain a huge number of cases, the global tendencies in the results shall be discussed first:
The localized coordinates of the class *in cage* except the bond stretches show a moderate localisation, the quantities *Tail length* range from 2.36 Åto 3.65 Å. However, the bond stretches are a special case and sometimes show a much better localization. The localized coordinates of the class *in chain* are perfectly localized with two exceptions (see below). This very good localization is also reflected in the quantities *Tail length*, about half a bond length for bond stretches, about one bond length for bond angles and about 1.7 bond lengths for torsions. For each *in chain* degree of freedom there is always exactly one localized coordinate to which this degree of freedom is assigned. For an ethylene group, for example, there are seven bond stretches, five bond angles at each carbon atom and two torsions.
The *at cage* bond stretches are perfectly localized like almost all members of the class *in chain*. The other types of coordinates of the class *at cage*, bond angles, out of plane angles and torsions, on the other hand, show variable behavior and are discussed in detail below.
For all three sample molecules (see above), the mode separated coordinates contain the "strong" bond stretches, i.e. the bond stretches for regular chemical bonding (the hydrogen bonds in the $(H_2O)_6$-cluster are excluded), as individual basis coordinates (see sect. 2.1.2 and sect. 2.1.3). The localization of this bond stretches is therefore skipped, since the "strong" bond stretches are perfectly localized already.

Mixture of different types for mode separated coordinates The mixture of basis coordinates with different types in localized mode separated coordinates only shows up in $(H_2O)_6$-cluster-chain. The reason for this is that there is one more group in the mode separated coordinates for $(H_2O)_6$-cluster-chain than for the other two examples. This additional group includes all basis coordinates that contain a hydrogen bond (see sect. 2.1.3) and thus contains basis coordinates of all types.
Another theoretical possibility for the mixing of different types in localized mode separated coordinates would be a mixture of bond angles and out of plane angles (see sect. 2.1.3). However, out of plane angles only exist in the example

molecule 3-spiro-rings-chain. Presumably because of the precisely planar geometry of the rings, no mixing of bond angles and out of plane angles is observed.

In chain bond angles with apex at the O-atom of Adamantane-chain The bond angle at the oxygen atom directly bound to the cage belongs to the *in-chain* coordinates. However, when locating mode separated coordinates with Boys' method or with Pipek-Mezey's method, this angle is not obtained as a perfectly localized coordinate. The bond angle at the oxygen is only about 80 % in its local group.
The cause is probably the very compact cage structure of adamantane, which means that there are many eigenvectors of $\mathbf{BB^t}$ with eigenvalue 0 (see sect. 2.1.1 and 2.1.3), which then also contain the bond angle on the oxygen atom (see eq. 22).

At cage angle coordinates for $(H_2O)_6$-cluster-chain For all variants of the localization one sees only one instead of two bond angles in the *at-cage* coordinates for the three atoms bound to the cage. These angular coordinates are the internal bond angles of two H_2O molecules and the H-O-C angle. When localizing mode separated coordinates, these three bond angles are localized perfectly. This is because the bond angles which do not contain a hydrogen bond form a linear independent group and are therefore already present as individual bond angles in the mode separated coordinates (see sect. 2.1.3).

At cage angle coordinates for Adamantane-chain When localizing mode separated coordinates, all 32 degrees of freedom for angular movements of the 16 atoms bound to the cage appear in the *at-cage* coordinates. The 31 bond angles are perfectly localized, whereas the one torsion is only moderately well localized.
When localizing delocalized coordinates, only 20 of the above 32 degrees of freedom are seen in the *at-cage* coordinates. All of these 20 degrees of freedom are represented by bond angles that are approximately 90% in their local group.

At cage angle coordinates for *3-spiro-rings-chain* All 24 degrees of freedom for angular motion of the 12 atoms directly bound to the three rings are reflected in the *at-cage* coordinates. This applies to both delocalized coordinates and mode separated coordinates. For the mode separated coordinates all 24 degrees of freedom are perfectly localized, for the delocalized coordinates this applies only to the 12 bond angles. On the other hand, out of plane angles or the torsion are only moderately well localized.

Table 4: Localization characteristics of the three classes of localized coordinates for $(H_2O)_6$-cluster-chain

Coord. type	No. **1**	% Type **1**	Tail length Å **1**	% Local **1**	No. **2**	% Type **2**	Tail length Å **2**	% Local **2**
				Boys Localization				
Stre in cage	18	93	2.27	88	18	94	1.78	89
Bend in cage	10	71	3.29	66	9	84	2.52	74
Tors in cage	5	86	2.98	87	9	89	2.86	93
Stre at cage	3	100	0.57	100	3	100	0.57	100
Bend at cage	3	87	2.83	74	3	100	1.02	100
Stre in chain	12	100	0.61	100	12	100	0.61	100
Bend in chain	20	100	1.51	100	20	100	1.51	100
Tors in chain	4	100	2.14	100	4	100	2.14	100
				Pipek-Mezey Localization				
Stre in cage	18	98	2.23	92	18	97	1.69	90
Bend in cage	9	85	3.65	63	11	84	2.78	65
Tors in cage	7	89	3.26	90	8	97	2.73	95
Stre at cage	3	100	0.57	100	3	100	0.57	100
Bend at cage	3	91	3.43	82	3	100	1.02	100
Stre in chain	12	100	0.61	100	12	100	0.61	100
Bend in chain	20	100	1.59	100	20	100	1.51	100
Tors in chain	4	100	2.14	100	4	100	2.14	100

Table 5: Localization characteristics of the three classes of localized coordinates for Adamantane-chain

Coord. type	No. **1**	% Type **1**	Tail length Å **1**	% Local **1**	No. **2**	% Type **2**	Tail length Å **2**	% Local **2**
Boys Localization								
Stre in cage	12	94	2.33	88	12	100	0.76	100
Bend in cage	12	94	2.36	73	12	100	3.37	79
Tors in cage	12	64	2.64	92	-	-	-	-
Stre at cage	16	100	0.57	100	16	100	0.57	100
Bend at cage	20	94	2.20	87	31	100	1.60	100
Tors at cage	-	-	-	-	1	100	3.34	62
Stre in chain	8	100	0.63	100	8	100	0.63	100
Bend in chain	11	100	1.50	100	11	100	1.83	98
Outp in chain	1	100	1.54	100	1	100	1.54	100
Tors in chain	4	100	2.14	100	4	100	2.14	100
Pipek-Mezey Localization								
Stre in cage	12	98	2.41	93	12	100	0.76	100
Bend in cage	12	99	2.81	72	12	100	3.00	77
Tors in cage	12	82	2.87	90	-	-	-	-
Stre at cage	16	100	0.57	100	16	100	0.57	100
Bend at cage	20	100	2.26	91	31	100	1.66	100
Tors at cage	-	-	-	-	1	100	3.34	62
Stre in chain	8	100	0.63	100	8	100	0.63	100
Bend in chain	11	100	1.53	100	11	100	1.81	97
Outp in chain	1	100	1.54	100	1	100	1.54	100
Tors in chain	4	100	2.14	100	4	100	2.14	100

2.3.4 Truncation of localized internal coordinates

For geometry optimizations or for the freezing of individual internal coordinates, the columns of the B-matrix of the internal coordinates do not have to be orthogonal to each other (see sect. 2.1.7). Therefore, one can delete the coefficients with smaller absolute values from the coefficient vectors of the localized coordinates as long as the set of localized coordinates does not become linearly dependent. A rough measure of the proximity to the linear dependency is the smallest eigenvalue of the $\mathbf{BB^t}$-matrix (see sect. 2.1.7).
The results are discussed below. First it is about the truncation of delocalized coordinates (see sect. 2.1.1) and then about the truncation of mode separated coordinates (see sect. 2.1.3) and of generalized natural coordinates (see sect. 2.1.4).

Truncation of delocalized coordinates First, the results for the Boys-localization of table 7 will be discussed here, the discussion of the results for the Pipek-Mezey localization of table 8 will follow later. In the first rows for each example molecule in table 7, the quotient of the smallest eigenvalue of $\mathbf{BB^t}$ and the reference value for different values of the quenching criterion for the coefficients is given. The reference value is the smallest eigenvalue of $\mathbf{BB^t}$ for the respective localized coordinates without deleted coefficients.
For each column of table 7, all coefficients with absolute values smaller than the respective criterion were deleted. For smaller values of the criterion, this quotient stays close to 1, presumably because the effect of deleting very many small coefficients cancel out in the statistical sense.
With increasing values of the threshold, the values of the quotients then fall. But up to a value of the criterion between 0.1 and 0.2 the localized coordinates with deleted coefficients, i.e. the *truncated coordinates*, remain far enough away from the linear dependence.
When localizing internal coordinates, the so called *orthogonality tails* arise, just like when localizing molecular orbitals. Obviously these tails get shorter as the value of the cutoff threshold increases. In the second and third rows of the sections for each example molecule in table 7, the average length of the tails and the maximum length of the tails are given for each value of the threshold. The length of the tail for a given localized coordinate is defined as follows: It is the greatest possible distance of a population $P_{\alpha,i}$ with a magnitude greater than 10^{-6} at any atom with index i from the centroid $\vec{Z}_\alpha$ of the given localized coordinate with index α (see equation 23).

Table 6: Localization characteristics of the three classes of localized coordinates for 3-spiro-rings-chain

Coord. type	No. **1**	% Type **1**	Tail length Å **1**	% Local **1**	No. **2**	% Type **2**	Tail length Å **2**	% Local **2**
				Boys Localization				
Stre in cage	15	88	2.98	84	15	100	0.77	100
Bend in cage	9	85	2.92	82	12	100	2.60	87
Outp in cage	3	59	3.59	70	-	-	-	-
Tors in cage	6	72	3.26	95	6	100	3.52	60
Stre at cage	12	100	0.56	100	12	100	0.56	100
Bend at cage	12	100	1.83	100	12	100	1.72	100
Outp at cage	12	53	3.38	72	12	100	1.63	100
Stre in chain	12	100	0.60	100	12	100	0.60	100
Bend in chain	20	100	1.46	100	20	100	1.46	100
Tors in chain	4	100	2.07	100	4	100	2.07	100
				Pipek-Mezey Localization				
Stre in cage	15	94	3.27	87	15	100	0.77	100
Bend in cage	9	93	3.26	71	12	100	2.65	86
Outp in cage	-	-	-	-	-	-	-	-
Tors in cage	9	90	3.22	91	6	100	3.53	60
Stre at cage	12	100	0.56	100	12	100	0.56	100
Bend at cage	12	100	1.71	100	12	100	1.71	100
Outp at cage	11	74	3.59	75	12	100	1.63	100
Tors at cage	1	86	2.75	87	-	-	-	-
Stre in chain	12	100	0.60	100	12	100	0.60	100
Bend in chain	20	100	1.61	100	20	100	1.61	100
Tors in chain	4	100	2.07	100	4	100	2.07	100

For a threshold of 0.1, the mean length of the tails is between 2.37 Åand 3.03 Å. For a threshold of 0.2, the mean length of the tails is only between 1.78 Åand 2.04 Å. These are good values considering that all example molecules are compact cages (see figures 1,5,6,8).
All intramolecular hydrogen bonds and all intermolecular hydrogen bonds in the $(H_2O)_6$-cluster and in Fructosyl-nistose-$(H_2O)_3$ have already been defined before the internal coordinates were generated.
The maximum lengths of the tails are significantly larger than the average lengths for all example molecules. This is to be expected, since there are far fewer degrees of freedom than redundant internal coordinates. Because of this, not all localised coordinates for cages can be perfectly localized (for definition see sect. 2.3.3).
In the fourth rows of the sections for each example in table 7 the percentages of coefficients with absolute values greater than 0 in the vectors of the localized coordinates are given. In the last two examples, with 114 atoms and 117 atoms respectively, the percentage is about 1.2 % for a criterion value of 0.1 and about 0.6 % for a criterion value of 0.2. This is not much for such relatively small example molecules.
The fifth row of each section in table 7 gives the percentage of atomic populations $P_{\alpha,i}$ (see eq.23) greater than zero in the B-matrices of the localized coordinates for each value of the cutoff criterion. The percentage of non zero populations in the last two examples is about 7 % for a threshold of 0.1 and about 5 % for a threshold of 0.2. This means, that a localized coordinate for a threshold of 0.1 extends over an average of 8 atoms, for a threshold of 0.2 it is an average of 5.7 atoms.
Table 8 has the same structure as table 7, but here the results for the Pipek-Mezey localization are given. With the cut-off criterions of 0.1 and 0.2, the quotients of the eigenvalues are slightly larger than for the Boys method. The average length of the tails at a threshold of 0.1 is about 10 % larger than for the Boys method. For the value of the criterion of 0.2 the average lengths of the tails are comparable. The difference between Boys and Pipek-Mezey is even more pronounced with the maximum lengths of the tails. For both values of the thresholds the maximum lengths of the tails are significantly larger.

Truncation of mode separated coordinates and generalized natural coordinates From the data in tables 9 and 10 one can see the shrinking *orthogonality tails* of localized mode separated coordinates and localized generalized natural coordinates with increasing cutoff parameters. This two types of coordinates would be important for practical

Table 7: Truncation of Boys-localized delocalized coordinates

Cutoff	0.002	0.005	0.01	0.02	0.05	0.1	0.20
$(H_2O)_6$-cluster							
Quotient	1.000	0.999	0.988	0.945	0.778	0.569	0.409
Mean tail Å	3.85	3.84	3.81	3.73	3.35	2.59	1.87
Max. tail Å	4.54	4.53	4.51	4.47	4.29	3.88	3.57
$\% > 0$ Coeff.	83	73	59	43	20	8.9	4.1
$\% > 0$ B-matrix	94	94	94	92	81	53	35
$C_{30}H_{16}$							
Quotient	0.999	0.997	0.987	0.973	0.847	0.642	0.553
Mean tail Å	5.77	4.95	4.50	4.02	3.27	3.03	2.04
Max. tail Å	10.42	10.29	8.43	7.39	5.16	4.73	3.63
$\% > 0$ Coeff.	18	14	11	8.4	5.0	3.1	1.7
$\% > 0$ B-matrix	55	45	40	33	25	21	14
$C_{30}H_{15}$-chain-corannulene							
Quotient	1.000	1.000	0.999	0.998	0.996	0.989	0.979
Mean tail Å	4.26	3.87	3.56	3.15	2.59	2.43	1.79
Max. tail Å	10.27	10.19	7.79	7.33	5.76	4.49	3.86
$\% > 0$ Coeff.	8.6	6.4	4.7	3.2	1.8	1.0	0.54
$\% > 0$ B-matrix	16	14	12	10	8.0	6.7	4.9
Fructosyl-nistose-$(H_2O)_3$							
Quotient	1.000	0.992	0.969	0.911	0.777	0.569	0.281
Mean tail Å	5.76	5.00	4.44	3.89	3.04	2.37	1.78
Max. tail Å	9.93	9.30	8.80	7.73	7.28	5.48	3.51
$\% > 0$ Coeff.	17	12	8.4	5.4	2.7	1.4	0.69
$\% > 0$ B-matrix	31	24	20	16	10	7.2	5.0

applications of my method. This is because the computing time for the localization is much smaller, especially for the generalized natural coordinates, than for delocalized coordinates (see Table 3).

Table 9 refers to the Boys-localization, the corresponding table 10 for the Pipek-Mezey localization will be discussed later. There are two example molecules, $C_{30}H_{15}$-chain-corannulene and Fructosyl-nistose-$(H_2O)_3$.

Table 9 is very similar to tables 7 and 8, the *quotient* and *max tail Å* entries are defined identically. In the middle of each block, *mean no. atoms* is the average number of atoms spanned by a localized coordinate. This quantity has the advantage that it can be directly transferred to larger example molecules.

With a value of the cut-off parameter of 0.1, a localized coordinate only extends over five atoms on average, with a value of 0.2 there are even fewer atoms. So Wilsons B-matrix would be for large molecules, e.g. Proteins, very sparsely populated. This can make the back-transformation from internal coordinates to cartesian coordinates much faster [2, 3, 12]. This would be important for the use of internal coordinates for large molecules, because this computational step is required in each cycle of a geometry optimization.

In my opinion, the optimal value for the cutoff parameter is again between 0.1 and 0.2. However, this cannot be seen so clearly from the data in Tables 7 and 8. This is because the delocalized coordinates are conceptually much simpler than the latter two types of internal coordinates (see sect. 2.1.1, 2.1.3 and 2.1.4).

The maximum length of the tails are slightly larger when localizing generalized natural coordinates compared to localizing delocalized coordinates. For mode separated coordinates this difference is even clearer. The explanation for this is probably that with the first two types of coordinates, fewer vectors are localized together than with the delocalized coordinates.

Table 10 is almost identical to Table 9, except that Pipek-Mezey's localization method was used for Table 10 instead of Boys' method. The results are very similar, only the maximum length of the tails are slightly larger.

Table 8: Truncation of Pipek-Mezey-localized delocalized coordinates

Cutoff	0.002	0.005	0.01	0.02	0.05	0.1	0.20
$(H_2O)_6$-cluster							
Quotient	1.000	1.001	1.003	1.002	0.929	0.648	0.317
Mean tail Å	3.78	3.77	3.76	3.69	3.43	2.78	1.73
Max. tail Å	4.57	4.57	4.55	4.53	4.30	4.28	4.14
$\% > 0$ Coeff.	81	71	58	42	20	8.5	3.4
$\% > 0$ B-matrix	94	94	94	93	85	58	31
$C_{30}H_{16}$							
Quotient	0.999	0.995	0.974	0.941	0.747	0.658	0.588
Mean tail Å	6.67	5.95	5.52	4.79	3.84	3.30	2.06
Max. tail Å	12.37	12.34	11.87	10.59	9.34	7.86	4.63
$\% > 0$ Coeff.	21	16	13	9.7	5.4	2.9	1.2
$\% > 0$ B-matrix	63	55	49	41	30	23	14
$C_{30}H_{15}$-chain-corannulene							
Quotient	1.000	1.000	1.000	0.999	0.995	0.997	0.974
Mean tail Å	4.54	4.23	3.92	3.57	3.00	2.59	1.76
Max. tail Å	12.24	12.24	11.86	11.53	9.28	7.93	4.83
$\% > 0$ Coeff.	8.8	6.8	5.1	3.5	1.8	1.0	0.47
$\% > 0$ B-matrix	17	16	14	12	9.4	7.4	4.7
Fructosyl-nistose-$(H_2O)_3$							
Quotient	1.000	0.999	0.987	0.965	0.846	0.795	0.683
Mean tail Å	6.18	5.55	5.01	4.38	3.35	2.56	1.69
Max. tail Å	11.30	10.78	9.27	9.00	7.78	7.34	4.78
$\% > 0$ Coeff.	19	13	9.3	5.8	2.5	1.1	0.53
$\% > 0$ B-matrix	35	28	23	18	12	7.3	4.6

2.4 Conclusions:

My localized internal coordinates have some resemblance to natural internal coordinates, especially for linear or branched chains (see sect. 2.3.3). Natural internal coordinates are considered the most appropriate internal coordinates for geometry optimizations [1]. Therefore geometry optimizations could be accelerated by relocalizing delocalized coordinates as much as possible with my methods.

Localized internal coordinates should also facilitate the *Freezing* of local degrees of freedom in molecules, because then only a few internal coordinates go into the freezing algorithm. This makes the freezing faster and probably more robust. Conversely, the localized internal coordinates also make it possible to optimize only a few local degrees of freedom in a large molecule.

I would recommend using delocalized coordinates based on contracted redundant internal coordinates (see sect. 2.1.2). This reduces the computation time for generating delocalized coordinates and also the computation time for the localization.

The generation of delocalized coordinates or of mode separated coordinates scales with the cube of the number of degrees of freedom of a molecule (see sect. 2.3.2). For large molecules, the degrees of freedom of as many atoms as possible must therefore be described by natural internal coordinates. This is because the generation of natural internal coordinates scales linearly.

With proteins one could e.g. describe the side groups of most amino acids, all other side chains and all terminal atoms by natural internal coordinates.

My relocalized delocalized coordinates only extend over about 8 atoms on average if one deletes all coefficients with absolute values less than or equal to 0.1 (see sect. 2.3.4) from the coefficient vectors. For my relocalized mode separated coordinates and for the relocalized generalized natural coordinates, this value is only about 5 atoms. Thus, the B-matrix (see sect. 2.1.1 and 2.1.7) for any molecule would be similarly sparse as for molecules whose degrees of freedom can be completely described by natural internal coordinates.

The back transformation of internal coordinates in cartesian coordinates and the transformation of gradients from cartesian coordinates into internal coordinates consume almost all of the computing time needed for the internal coordinates

Table 9: Truncation of Boys-localized **MS** and **GN** coordinates

Cutoff	0.0	0.005	0.01	0.02	0.05	0.10	0.20
$C_{30}H_{15}$-chain-corannulene, mode separated coordinates							
Quotient	1.000	0.997	0.982	0.965	0.950	0.948	0.791
Mean no. atoms	70.7	8.62	7.71	6.82	5.72	5.00	4.39
Max. tail Å	25.91	8.20	7.35	6.74	5.51	5.62	4.63
$C_{30}H_{15}$-chain-corannulene, generalized natural coordinates							
Quotient	1.000	0.995	0.985	0.962	0.922	0.931	0.866
Mean no. atoms	21.51	6.71	6.26	5.81	5.07	4.54	4.06
Max. tail Å	25.42	8.14	6.91	6.26	5.44	4.86	3.94
Fructosyl-nistose-$(H_2O)_3$, mode separated coordinates							
Quotient	1.000	1.011	1.013	0.979	0.824	0.716	0.730
Mean no. atoms	70.15	12.83	10.12	8.12	5.98	5.12	4.56
Max. tail Å	16.13	10.56	10.17	10.22	10.60	7.57	5.22
Fructosyl-nistose-$(H_2O)_3$, generalized natural coordinates							
Quotient	1.000	0.989	0.981	0.931	0.926	0.909	0.652
Mean no. atoms	13.43	8.32	7.22	6.22	5.02	4.36	3.95
Max. tail Å	10.32	9.95	9.62	9.33	8.63	5.47	5.32

Table 10: Truncation of Pipek-Mezey-localized **MS** and **GN** coordinates

Cutoff	0.0	0.005	0.01	0.02	0.05	0.10	0.20
$C_{30}H_{15}$-chain-corannulene, mode separated coordinates							
Quotient	1.000	0.994	0.986	0.975	0.932	0.963	0.949
Mean no. atoms	70.7	8.77	7.91	7.23	6.01	5.20	4.40
Max. tail Å	26.78	9.49	8.65	6.93	5.67	5.62	4.75
$C_{30}H_{15}$-chain-corannulene, generalized natural coordinates							
Quotient	1.000	0.997	0.999	0.999	0.983	1.103	1.193
Mean no. atoms	21.51	7.14	6.63	6.11	5.42	4.72	4.14
Max. tail Å	26.33	9.86	8.43	7.71	6.25	5.52	4.10
Fructosyl-nistose-$(H_2O)_3$, mode separated coordinates							
Quotient	1.000	1.004	1.017	0.989	1.034	1.187	1.161
Mean no. atoms	70.15	11.69	9.56	7.62	5.87	5.09	4.57
Max. tail Å	16.34	12.59	12.19	10.00	10.30	7.93	7.57
Fructosyl-nistose-$(H_2O)_3$, generalized natural coordinates							
Quotient	1.000	0.992	0.997	1.016	1.052	0.974	0.857
Mean no. atoms	13.43	8.17	7.08	6.11	5.04	4.45	3.99
Max. tail Å	10.31	9.90	9.54	9.37	9.31	6.65	5.29

in the iteration cycles of geometry optimizations or of transition state searches. By Németh, Coulaud, Monard and Ángyán [2] and by Paizs, Fogarasi, Baker, Suhai and Pulay [3, 12], methods are described which enable the above transformations with linear scaling. These methods could be used with my truncated internal coordinates for any molecule, because the B-matrix for truncated internal coordinates is very sparse (see above).
Instead of the Jacobi rotations (see sect. 2.1.5), one could also use more modern localization algorithms, to improve the convergence of the localization (see sect. 2.3.1). This would be e.g. the *Trust region minimization* method of Høvik, Jansik and Jørgensen [13, 14].
Using my simplified versions of the Boys algorithm and of the Pipek-Mezey algorithm the conzept of **localization** can be generalized to a set of n vectors in an Euklidian and real valued vector space of dimension $\mathbf{m}$. The basis vectors

must be orthonormal, with $\mathbf{n} \geq \mathbf{2}$ and $\mathbf{m} \geq \mathbf{2}$ and $\mathbf{m} \geq \mathbf{n}$. But another condition is that you can assign a location vector in an Euklidian and real valued l-dimensional vector space with orthonormal basis vectors, $\mathbf{l} \geq \mathbf{1}$, to each basis vector of the n vectors mentioned above in a canonical way.
Descriptively, localization means that for each vector of the set mentioned above the basis vectors with larger absolute values of the coefficients are only distributed over as small an area in the l-dimensional space as possible.
Because of the general applicability of the simplified localization methods, there could be another application in the natural sciences or in mathematics.
Jacob, Panek and Reiher have localized normal vibrations, the algorithms are analogous to the methods of Boys and Pipek-Mezey [15, 16]. However, these articles do not consider a more general application of localization.

3 New intermolecular coordinates for supermolecules in Turbomole:

3.1 Theory:

Weak bonds have significantly smaller binding energies than ordinary chemical bonds and they are always needed when a molecule consists of several non-connected parts. In this case, the molecule must be made coherent by redefining weak bonds, because otherwise not all degrees of freedom of the molecule can be described by internal coordinates.
In the previous version of Turbomole, the additional bonds between two subunits were defined **first**. This redefinition of bonds was iterated until the molecule was recognized as fully connected. After that, the procedure was essentially the same as for ordinary molecules consisting only of one piece.
However, adding a new bond can result in unusual ligand arrangements at the two atoms involved, which the internal coordinate generator considers "unchemical" and therefore rejects. Not enough local internal coordinates were then defined for the atoms in question. This problem occurs particularly often when several new bonds are formed on an atom or when the two subunits to be connected are far apart.
In order to avoid these problems, internal coordinates for each of the isolated parts of the molecule are first defined separately in my improved version. **Only then** is the molecule made cohesive by additional bonds.

Hydrogen Bonds The most important and at the same time the strongest weak bonds are the hydrogen bonds. For this reason, they are given a preferential treatment in Turbomole. The first attempt is to completely connect a molecule solely by redefining hydrogen bonds. Only when this is not possible is a second general class of weak bonds used (see below).
An hydrogen bond is defined if:

1. The **H**-atom and the partner atom **X** must belong to **different** subunits of the molecule.
2. The partner atom **X** must be a strongly electronegative atom, i.e. **F**, **O**, **N** or **Cl**.
3. Das **H**-Atom must be bound to strongly electronegative atom, i.e. **F**, **O**, **N** oder **Cl**.
4. For both the **H**-Atom and the partner atom **X**, the other atom must be the **nearest** atom of the opposite subunit of the molecule.
5. The distance between **H** and **X** must not be larger than **6.0** Bohr.
6. The angle **Y-H...X** with apex at **H** must be greater than **125** degrees.
7. An **H**-Atom can only form **one** hydrogen bond.

General weak Bonds At most **one** weak bond can be defined for each pair of subunits in a molecule, and only if the subunits involved are sufficiently close (see below). However, symmetry images of a new weak bond are of course also defined. There are **two** criteria for the formation of a weak bond: One is the bond distance and the other is the distance of the center of the weak bond to the common center of mass of the subunits to which the atoms to be bonded belong. A weighted sum of the two criteria is minimized and they are both weighted with a factor of **0.5**, but this weighting factor can be changed by user input.
If the symmetry group of a molecule is more than C1, then the coordinates of the bond center and the coordinates of the center of mass are symmetrized, i.e. they are projected onto the total symmetric representation.
The redefinition of weak bonds is iterated until the molecule is completely connected.
Two subunits are considered sufficiently close to form a new bond if: The shortest distance between the two subunits must not be greater than 1.3 times *Rminmin*. *Rminmin* is the shortest distance from any of the two subunits to any second subunit.

Figure 1: $(H_2O)_6$-cluster

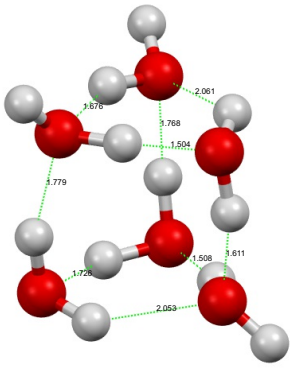


Figure 2: $(H_2O)_6$-cluster-chain

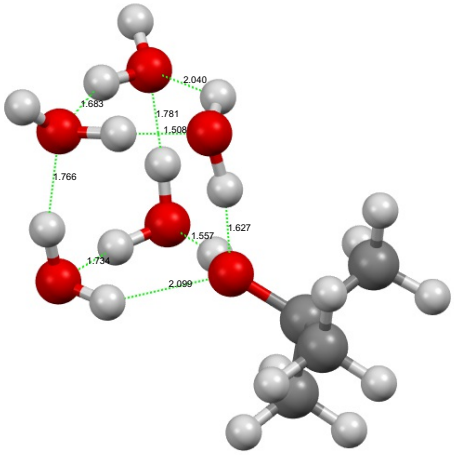


Figure 3: 3-spiro-rings-chain

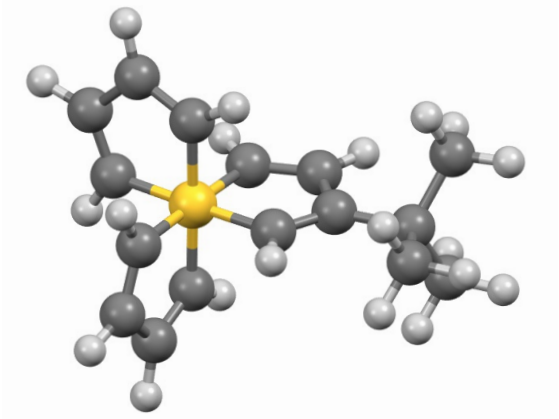


Figure 4: Adamantane-chain

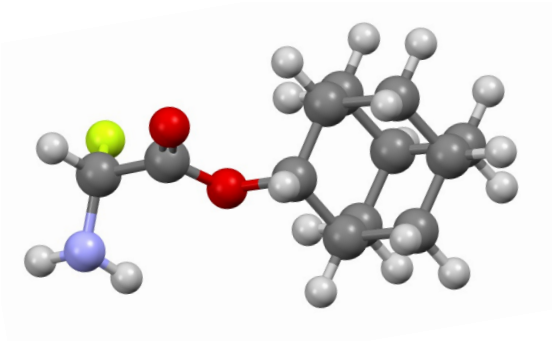


Figure 5: $C_{30}H_{16}$

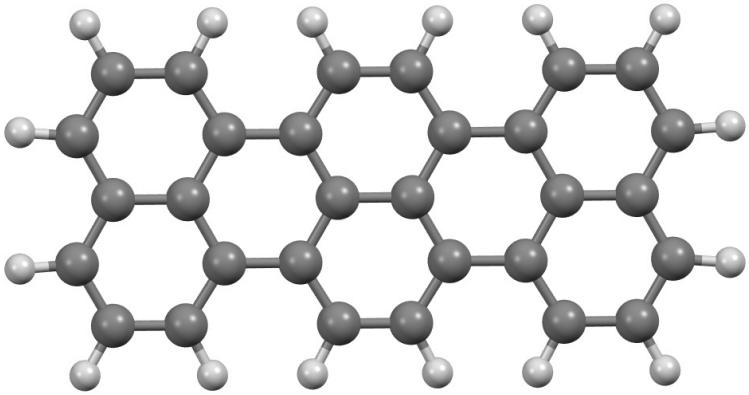


Figure 6: $C_{30}H_{15}$-chain-corannulene

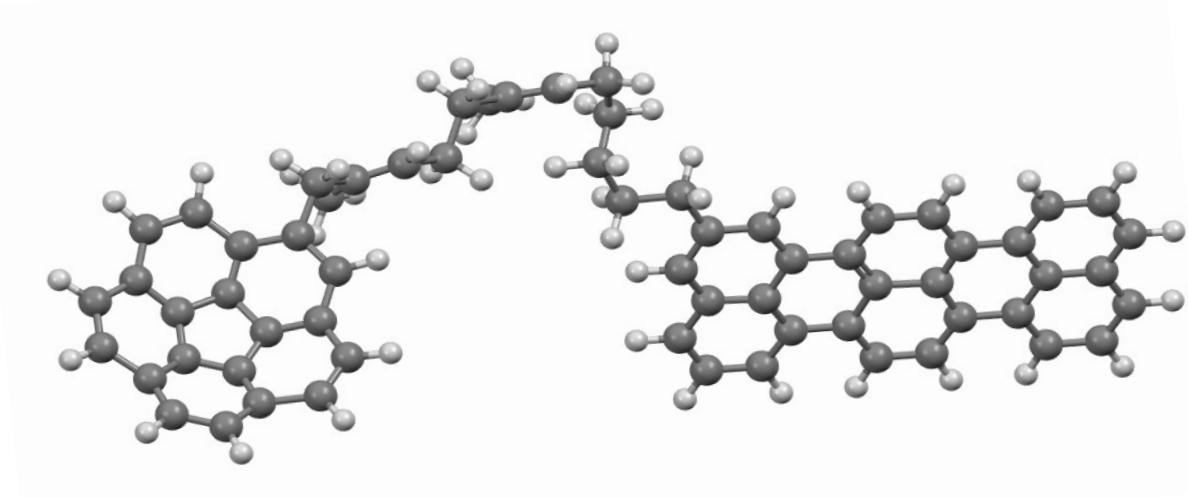


Figure 7: $C_{30}H_{13}$-prop-phen-naph

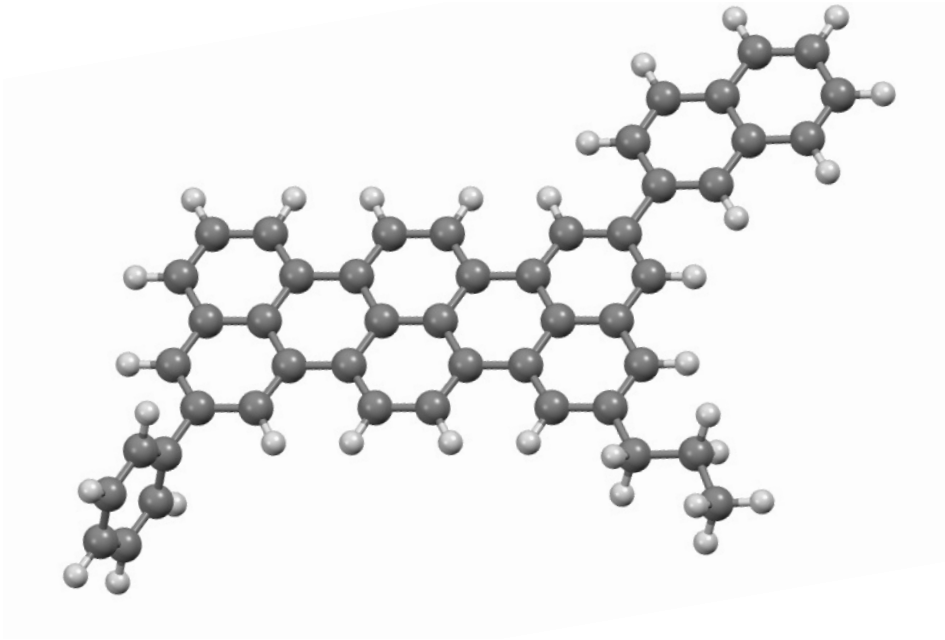
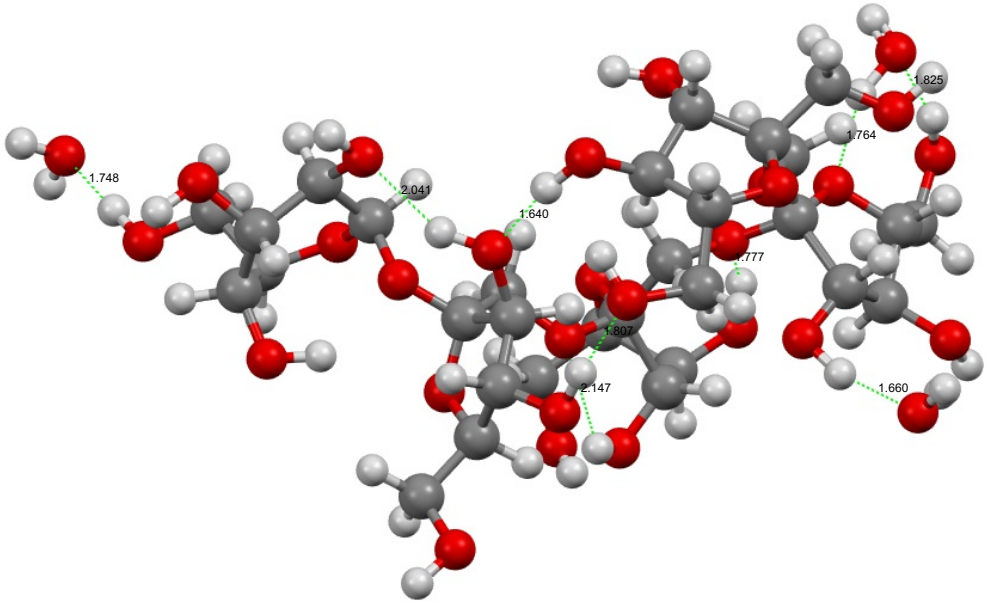


Figure 8: Fructosyl-nistose-$(H_2O)_3$



Generation of the intermolecular internal coordinates As a first step, a generally redundant set of so called *Link coordinates* is created. All link coordinates are primitive internal coordinates (see section 2.1.2). A link coordinate extends over at least two subunits of the supermolecule and it must contain at least one weak bond. That is, a weak bond has been redefined between at least two of the atoms used to define that link coordinate.
In order to get rid of the redundancy mentioned above, the procedure is then as follows: The algorithm for generating the mode separated coordinates is used (see sect. 2.1.3). However, in this application of the algorithm there are only two blocks, the first consists of the internal coordinates of the isolated sub-molecules and the second block contains all the link coordinates. Because the first block is already fixed and contains exactly one internal coordinate for each degree of freedom of the sub-molecules, the method is finished in the first iteration (cf. sect. 2.1.3).
The eigenvectors of the first block are of course discarded because there are already internal coordinates for the sub-molecules. The eigenvectors of the second block define exactly one internal coordinate for each degree of freedom of the intermolecular movements. These are the movements of the sub-molecules against each other.
A completely different approach to the description of the intermolecular degrees of freedom was developed by Wang and Song [17]. Some of the formalism was taken from Coutsias, Seok and Dill. [18].

3.2 Technical Details:

My new intermolecular internal coordinates were implemented into the Turbomole version of January 2018, which is the reference version. All source files were compiled with the GNU Fortran 95 compiler. The new coordinates are part of the current version of Turbomole and can be switched on by user input.
My example molecules (see figures 1,9,10,11,12 and 13) were created as described in section 2.2. However, the structures were distorted manually after their creation in order to increase the average number of cycles in the geometry optimizations. All informations about the test molecules is available via my email address (see front page).
The geometry optimizations for table 11 have all been done with the Turbomole defaults: This is the RIDFT method with def-SV(P) basis sets for each atom and with the BP86 density functional (see Turbomole manual Version 7.4 and references therein). *Jobex* was used as the driver script for the optimization cycles, and the geometry relaxations were carried out with the program *Statpt*.
The *generalized natural coordinates* (see section 2.1.4) were always used as internal coordinates for the reference calculations and as internal coordinates for the individual sub-molecules of the supermolecules (see section 3.1).

3.3 Results and discussion:

In test calculations on a few hundred examples, no errors were found for the new intermolecular coordinates. The reference version, on the other hand, crashed in some test examples. As expected, this happened mainly for complexes with large intermolecular distances, but also for some other multicomponent molecules.
In table 11 there are six molecules on which the geometry optimization was tested with the new coordinates. For BN-Graphene much fewer optimization cycles are needed than for the geometry optimization with the reference version. For the other molecules, the number of cycles is comparable. For BN-Graphene many additional bonds between the Graphene-layers are defined in the calculation with the reference version. This probably leads to strong couplings between the intramolecular degrees of freedom and the intermolecular degrees of freedom and therefore to a significant slowdown in geometry optimization.
There are strong improvements in computing time for the definition of the internal coordinates. This is especially true when the largest subunit of a molecule contains not much more than half of the atoms.
$\mathbf{Al_{147}}$ is a highly coordinated cluster and therefore has many redundant internal coordinates (2736) and 435 degrees of freedom. The definition of internal coordinates for two separate $\mathbf{Al_{147}}$-clusters with the defaults of the reference version takes 389 seconds, with my new intermolecular coordinates it is only 105 seconds.
Because the linear algebra takes the lion's share of the computing time, you see approximately the theoretical factor of 4. In other test examples, however, the additional effort for describing the movement of the subunits in relation to one another could have a higher share.

3.4 Conclusions:

My new intermolecular coordinates make the definition of internal coordinates for supermolecules more robust and also faster. When applied to large molecules, e.g. on complexes of multiple globular proteins, this could make the definition of internal coordinates workable for such cases.
In some examples, the number of cycles for geometry optimizations is also greatly reduced (see sect. 3.3).

Figure 9: Benzene-Cr-Benzene

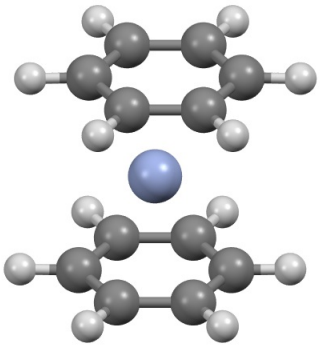


Figure 10: BN-Graphene

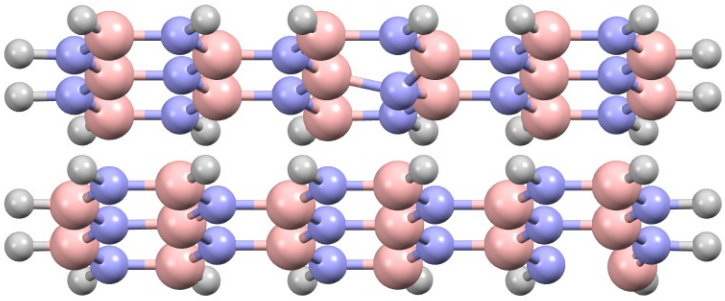


Figure 11: Guanin-Cytosin

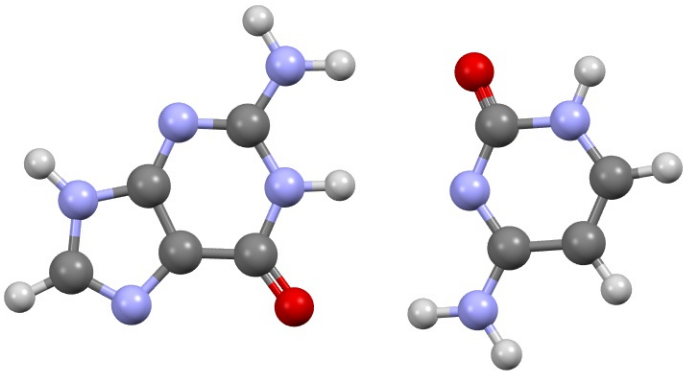


Table 11: Number of cycles required for the Geometry Optimization

Molecule	Reference Version	new linking coords.
Benzene-Cr-Benzene	22	25
BN-Graphene	115	28
Guanine-Cytosine	37	33
$(H_2O)_6$	24	19
Naph$(OH)_2$-NNnaph$(OH)_2$	52	44
t-Butanol-Dimer	38	33

References

[1] Peter Pulay, Géza Fogarasi, Frank Pang, and James E. Boggs. Systematic ab initio Gradient Calculation of Molecular Geometries, Force Constants, and Dipole Moment Derivatives. *Journal of the American Chemical Society:*, volume 110. , pages 2550–2560. , 1979.

[2] Károly Németh, Olivier Coulaud, Gérald Monard, and János G. Ángyán. Linear scaling algorithm for the coordinate transformation problem of molecular geometry optimization. *The Journal of Chemical Physics:*, volume 113. , pages 5598–5603. , 2000.

[3] Béla Paizs, Géza Fogarasi, and Peter Pulay. An efficient direct method for geometry optimization of large molecules in internal coordinates. *The Journal of Chemical Physics:*, volume 109. , pages 6571–6576. , 1998.

[4] Malte von Arnim, and Reinhart Ahlrichs. Geometry optimization in generalized natural coordinates. *The Journal of Chemical Physics:*, volume 111. , pages 9183–9190. , 1999.

[5] Jon Baker, Alain Kessi and Bernard Delley. The generation and use of delocalized internal coordinates in geometry optimization. *The Journal of Chemical Physics:*, volume 105. , pages 192–212. , 1996.

[6] Susi Lehtola, and Hannes Jonsson. Unitary Optimization of Localized Molecular Orbitals. *Journal of Chemical Theory and Computation:*, volume 9. , pages 5365–5372. , 2013.

[7] Peter Pulay, and Géza Fogarasi. Geometry optimization in redundant internal coordinates. *The Journal of Chemical Physics:*, volume 96. , pages 2856–2860. , 1992.

[8] Károly Németh, Matt Challacombe, and Michel Van Veenendaal. The choice of internal coordinates in complex chemical systems. *Journal of computational chemistry:*, volume 31. , pages 2078–2086. , 2010.

[9] S.F. Boys. Construction of Some molecular Orbitals to Be Approximately Invariant for Changes from One Molecule to Another. *Reviews of Modern Physics:*, volume 32. , pages 296–299. , 1960.

[10] János Pipek and Paul G. Mezey. A fast intrinsic localization procedure applicable for *an initio* and semiempirical linear combination of atomic orbital wave functions. *The Journal of Chemical Physics:*, volume 90. , pages 4916–4926. , 1989.

[11] E. B. Wilson, Jr., J. C. Decius, and P. C. Cross. *Molecular Vibrations:* (Mc Graw-Hill, New York, 1955).

[12] Béla Paizs, John Baker, Sandor Suhai, and Peter Pulay. Geometry optimization of large biomolecules in redundant internal coordinates. *The Journal of Chemical Physics:*, volume 113. , pages 6566–6572. , 2000.

[13] Ida-Marie Høvik, Branislav Jansik, and Poul Jørgensen. Trust Region Minimization of Orbital Localization Functions. *Journal of Chemical Theory and Computation:*, volume 8. , pages 3137–3146. , 2012.

[14] Ida-Marie Høvik, Branislav Jansik, and Poul Jørgensen. Pipek-Mezey Localization of Occupied and Virtual Orbitals. *Journal of Computational Chemistry:*, volume 34. , pages 1456–1462. , 2013.

[15] Christoph R. Jacob and Markus Reiher. Localizing normal modes in large molecules. *The Journal of Chemical Physics:*, volume 130. , page 084106. , 2009.

[16] Pawel T. Panek and Christoph R. Jacob. On the benefits of localized modes in anharmonic vibrational calculations for small molecules. *The Journal of Chemical Physics:*, volume 144. , page 164111. , 2016.

[17] Lee-Ping Wang, and Chenchen Song. Geometry optimization made simple with translation and rotation coordinates. *The Journal of Chemical Physics:*, volume 144. , page 214108. , 2016.

[18] Evangelos A. Coutsias, Chaok Seok, and Ken A. Dill. Using quaternions to calculate RMSD. *Journal of Computational Chemistry:*, volume 25. , pages 1849–1857. , 2004.

Figure 12: Naph$(OH)_2$-NNnaph$(OH)_2$

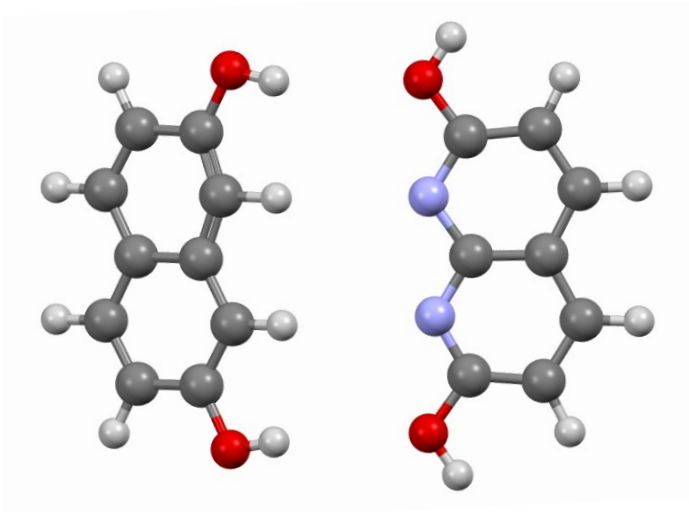


Figure 13: t-Butanol-Dimer