

Comparison of the Boys and Pipek–Mezey Localizations in the Local Correlation Approach and Automatic Virtual Basis Selection

James W. Boughton and Peter Pulay*

Department of Chemistry, University of Arkansas, Fayetteville, Arkansas 72701

Received 31 August 1992; accepted 30 December 1992

Our implementation of Pipek–Mezey population localization is described. It is compared with other localization schemes and its use in the framework of the local correlation method is discussed. For such use, this localization is shown to be clearly superior to Boys in the case of physically well-localized systems. Our current algorithm for selection of local virtual spaces is also described. © 1993 by John Wiley & Sons, Inc.

INTRODUCTION

Even with the development of high-speed supercomputers, there will continue to be, for the foreseeable future, systems of chemical interest for which conventional *ab initio* calculations incorporating electron correlation effects will be prohibitively expensive. More efficient high-level methods must continuously be sought. The local correlation (LC) method^{1–8} shows substantial reduction in computation time with essentially no change in quality of the results. The method is also well defined, unlike many other approximation schemes. Until recently, we used only Boys localization⁹ in our approach. Our exploration of the Pipek–Mezey (P–M) localization (PML)¹⁰ as an alternative has shown it to have distinct advantages in the local correlation method.

The LC method uses localized molecular orbitals (LMOs) in the internal (strongly occupied) space. This leads to a number of advantages:

1. The correlation energy can be decomposed into physically meaningful contributions.
2. Correlation contributions between spatially distant localized orbitals can be omitted or treated at a lower theoretical level.^{5,6}
3. The virtual space to describe correlation for a given electron pair can be restricted to a smaller set of functions with essentially no loss in the correlation energy. (Analysis of the CI problem shows that essentially all of the correlation energy arises from substitutions in which the virtual spaces are localized on the same atoms as the internal LMOs for which they are substituted. *In the LC method, we therefore define a local virtual space for each localized internal orbital.*)

4. Implementation of point 3 inherently reduces basis set superposition error (BSSE).^{2–6,11}

Points 2 and 3, but especially 3, are responsible for the computational savings in the LC method.

For an internal orbital describing a two-center bond, the local virtual space consists of the AOs located on these two centers, with the occupied space projected out. For a lone pair, the AOs on its one center are used. One option that may be justified is to use only the valence AOs when constructing local virtual spaces. Most localized valence orbitals can clearly be categorized as bicentric bonds or lone pairs. For orbitals that are genuinely delocalized, more centers will have to be included. We call genuinely delocalized orbitals those that resist orthogonal localization techniques, i.e., those that even after localization still have large populations on more than two atomic centers. As described in the next section, the automatic virtual space selection indicates the extent of delocalization. These results agree well with chemical intuition.

The virtual space for the most important double substitutions is defined as the *union* of the local spaces for the two internal MOs. For other fine points of the method, such as the treatment of the singles, see refs. 1–8.

We emphasize the importance of truncating the virtual space individually for each localized MO. This has two advantages: It is well defined and causes only a minor truncation of the correlation energy. This advantage is not shared by other truncation schemes. For instance, the method advocated by Bartlett and Adamowicz tries to restrict the correlation space globally [called the optimized virtual orbital space method (OVOS) by Bartlett^{12,13} and the first-order correlation orbital space method (FOCO) by Adamowicz¹⁴]. This has two disadvantages: First,

*Author to whom all correspondence should be addressed.

there is no unique way to partition the virtual space. Second, if a well-balanced basis set is used there should be no basis functions (with the possible exception of very diffuse ones) that play no role in the correlation at all.

An obvious disadvantage of individual correlation spaces is that it is difficult to guarantee that the same truncation error is made, say, when comparing different isomers. Our experience is that this is not a serious problem. For instance, in our recent work on the six tautomers of uracil (to be published) done at the MP4(SDQ)/6-31G** level the LC method recovered at least 98.4% of the conventionally calculated correlation energies in every case. Moreover, the truncation error diminishes as the basis set quality is improved, so this problem probably will be less important in future high-quality calculations.

AUTOMATIC VIRTUAL SPACE SELECTION

Previously in our method, definition of the correlation space has been carried out manually, providing ample opportunity for mistakes. Program TX90 now handles this automatically.

For a given LMO, one, two, three, or four atoms are selected whose combined basis functions best span the LMO space (it has been found that four atoms are usually enough to adequately span a genuinely delocalized orbital). The spanning criterion is based upon a least-squares fit of the functions on the selected atoms to the LMO. We require a selection to span its corresponding LMO by at least 98%, judged by the least-squares residual. To assure an atomic selection of the appropriate symmetry, the electron density within the LMO is used to prioritize the atoms for selection. For each LMO case, the atoms are ranked according to a Mulliken-type population analysis within that LMO space ("gross Mulliken population"). The atomic selection is accumulated in the order of greater to less electron density (absolute value) and the spanning criterion (least-squares fit) is only applied as a test for ending the accumulation. This atom selection strategy seems to work well.

For a given LMO, the least-squares residual (which by our spanning criterion will have to be less than 0.02) for a *proposed selection* of atoms is

$$f(\mathbf{a}) = \min \left[\int (\varphi_i - \varphi'_i)^2 d\tau \right]$$

where

$$\varphi_i = \sum_{\lambda=1}^m C_{i\lambda} \chi_{\lambda} \quad \text{and} \quad \varphi'_i = \sum_{\alpha \in A} a_{i\alpha} \chi_{\alpha}$$

$C_{i\lambda}$ = the LCAO coefficients of MO i

A = the subspace of λ s resident only on selected atoms (1)

Proceeding, we must solve for all $\alpha_{i\beta}$

$$\frac{\partial f(\mathbf{a})}{\partial a_{i\alpha}} = -2 \sum_{\lambda=1}^m C_{i\lambda} S_{\lambda\alpha} + 2 \sum_{\beta \in A} a_{i\beta} S_{\alpha\beta} = 0 \quad (2)$$

or in matrix notation

$$\mathbf{S}_A \mathbf{a}_i = \mathbf{v}_i = \mathbf{S}_v \mathbf{c}_i \quad (3)$$

where \mathbf{S}_A = the overlap matrix of collected elements whose rows and columns span the functional space A and \mathbf{S}_v = the overlap matrix whose columns span the space A but whose rows span the molecular basis. The least-squares residual of the proposed atom selection for MO i is then

$$f(\mathbf{a}) = 1 - \mathbf{a}_i^T \mathbf{v}_i \quad (4)$$

COMPARISON OF LOCALIZATION SCHEMES

The most widely used localization schemes are due to Boys⁹ and Edmiston and Ruedenberg (E-R).¹⁵⁻¹⁷ Both of these methods have localization criteria that can be physically rationalized. The Boys method maximizes the separation of orbital centroids while the E-R method maximizes the orbital self-repulsion energies. For double bonds, Boys gives a set of "τ" or "banana bonds," while the E-R method gives conventional σ and π bonds. Both of these methods are compared in ref. 18.

The Pipek-Mezey method does not localize in terms of Coulombic properties as do the above two methods, but its orbitals are similar to the E-R orbitals, particularly in their σ - π separation. This σ - π separation is an inherent mathematical property of the PML method.¹⁰ The PML method will be seen to have other theoretical advantages as well. As the Pipek-Mezey localization is recent, we here summarize its main features.

Pipek has presented a well-defined general measure of localization (delocalization)¹⁹:

$$d_i = \left[\sum_{A=1}^n (Q_A^i)^2 \right]^{-1} = \left[\sum_{A=1}^n \left(\sum_{\mu \in A} \sum_{\nu=1}^m C_{i\mu} C_{i\nu} S_{\mu\nu} \right)^2 \right]^{-1} \quad (5)$$

where Q_A^i = the gross Mulliken population of orbital i on atom A , $C_{i\mu}$ = MO coefficient of AO μ in MO i , $S_{\mu\nu}$ = the AO overlap integral, and n = the number of atoms. It is easy to see that this generally gives the number of atoms the MO is localized (delocalized) on. To illustrate, for a *strictly* bicentric bond

$$Q_1 = Q_2 = 1/2, d = (1/4 + 1/4)^{-1} = 2 \quad (6)$$

Physical interpretations can be given to the deviations, + or -, from an integral valued d_i , i.e., the extent of MO delocalization tails and bond polarization, respectively.¹⁹

A mean localization quantity is defined in ref. 10 as

$$D^{-1} = \frac{1}{N} \sum_{i=1}^N d_i^{-1} \quad (7)$$

but is taken more simply—for the global localization criterion—as

$$P = \sum_{i=1}^N \sum_{A=1}^n (Q_A^i)^2 = \text{maximum} \quad (8)$$

where N = the number of MOs. This global localization criterion essentially tends to minimize, for as many orbitals in the molecule as possible, the number of atoms that each orbital is localized on. Because of the LC virtual space truncation scheme, this is a desirable property.

In our implementation of PML, the transformation to localized RHF orbitals is realized by a converging iteration of 2×2 rotations wherein the angle of rotation (γ) for each iteration is found by:

$$\begin{aligned} \gamma_{st} &= \text{SIGN}(B_{st}) \frac{1}{4} \cos^{-1} \left(-\frac{A_{st}}{\sqrt{A_{st}^2 + B_{st}^2}} \right); \\ -\frac{\pi}{4} &\leq \gamma_{st} \leq \frac{\pi}{4} \\ A_{st} &= \sum_{A=1}^n \left\{ (Q_A^s)^2 - \frac{1}{4} [Q_A^s - Q_A^t]^2 \right\} \\ B_{st} &= \sum_{A=1}^n Q_A^s [Q_A^s - Q_A^t] \\ Q_A^{st} &= \frac{1}{2} \sum_{\mu \in A} \sum_{\nu=1}^m [C_{s\nu} C_{t\mu} + C_{s\mu} C_{t\nu}] S_{\mu\nu} \end{aligned} \quad (9)$$

The algorithmic complexity for PML is less than for the E-R method, and although on the order of Boys the actual convergence is faster than for Boys. Characteristics of the above three methods are summarized in Table I.^{10,18} See ref. 10 for a graphic comparison of Boys and E-R orbitals with Pipek orbitals for a number of molecules.

As compared to PML, Boys localization has convergence problems with certain types of physical systems. A case in point is fluoroethane. The Boys localization criterion hardly changes if the three fluorine lone pairs are rotated simultaneously, in a manner analogous to the rotation of hydrogens in a methyl torsion. Optimization on such shallow surfaces is a notoriously difficult process and may not

even be unique. This difficulty arises because Boys tries to *physically* localize in terms of *spatial extension*. As indicated in eqs. (5)–(8), the PML method is not so oriented. It only tries to make a *population* localization in terms of *atomic centers*. Pipek–Mezey localization therefore gives us a unique and more rapid localization in such cases as fluoroethane, as well as generally. Because both Boys and P–M localizations are carried out using 2×2 rotations, a count of these rotations will indicate the relative rates of convergence for these two methods. A comparison of such counts for the molecules of this work can be found in Table II.

Symmetry breaking is another difficulty that can show up with Boys localization. This is clearly seen in the case of planar hydrazine when all occupied MOs are included in the localization. The N–H and N–N bond (LMO) centroids are out of plane one way by 0.05 bohrs and 0.02 bohrs respectively, while the lone pairs are out of plane the other way by 0.12 bohrs (geometry optimized at 6-31G** with offset forces). The corresponding PML is perfectly symmetrical. With such symmetry breaking, for some symmetrical molecules we must forego the computational advantages that could be exploited in correlation calculation. The Boys localization can be, of course, symmetry constrained. In planar hydrazine, for instance, one can localize σ and π orbitals separately. However, localization changes drastically for the slightest deviation from planarity, making such a method unsuitable for potential surfaces.

There is a distinction to be made between PML and Boys localization with regard to localized and delocalized systems. PML may be slightly disadvantageous for genuinely delocalized systems. This is because its localization criterion (8) will allow some of the extended orbitals not to localize so much as the others to gain a better *global* result. But, when we truncate the local spaces of these delocalized orbitals they are then not so well represented. Of course, for physically well localized systems this problem does not arise and PML is then advantageous where the local space truncation is concerned. Conversely, the lack of σ – π separation in Boys localization allows delocalized π orbitals to be somewhat better localized than with PML due to mixing with σ orbitals, giving the τ banana bonds. This is at the expense of less extended orbitals, making them a little more delocalized. But, this is still acceptable as they are still well represented with a truncation

Table I. Characteristics of different localization schemes.

Localization method	Optimum localization criteria	Algorithmic complexity
Edmiston–Ruedenberg	Maximize the orbital self-repulsion energies	N^5
Boys	Maximize distances between orbital centroids	N^3
Pipek–Mezey	Minimize the number of atoms orbital is localized on	N^3

Table II. Convergence characteristics and conventional MP2, MP3, and MP4(SDQ) correlation energies and local correlation energies (as percentages of the conventional results) for the Boys and Pipek-Mezey localizations.

	No. of passes	2 × 2 rots.	$E(\text{MP2})$, local %	$E(\text{MP3})$, local %	$E(\text{MP4(SDQ)})$, local %	Total energy, 3,4-center MOs
Diborane			-.218089	-.250262	-.255151	-53.073783
Boys	7	105	98.558	98.611	98.632	2 0
P-M	7	83	98.562	98.614	98.635	2 0
Furan			-.710508	-.728564	-.737011	-229.368440
Boys	24	1872	98.922	99.013	98.995	0 5
P-M	8	310	98.674	98.757	98.755	0 3
Benzene			-.792343	-.823826	-.829071	-231.542116
Boys	25	2625	98.671	98.781	98.794	0 6
P-M	10	430	98.363	98.507	98.519	0 3
Fluoroethane			-.463216	-.487435	-.495038	-178.580307
Boys	762	3776	98.534	98.598	98.614	0 0
P-M	46	923	98.540	98.6	98.616	0 0
Dimethyl ether			-.478361	-.506356	-.513764	-154.587003
Boys	13	585	98.332	98.472	98.444	0 0
P-M	8	274	98.499	98.593	98.584	0 0
Oxirane			-.462131	-.480149	-.487193	-153.360259
Boys	14	504	98.448	98.567	98.548	0 0
P-M	7	188	98.634	98.701	98.702	0 0
Methylamine			-.328810	-.352752	-.356612	-95.577941
Boys	10	210	98.333	98.426	98.454	0 0
P-M	7	124	98.339	98.43	98.457	0 0
Oxalic acid			-.965053	-.960010	-.979971	-377.361112
Boys	83	4726	98.59	98.794	98.640	0 1
P-M	9	574	99.065	99.136	99.084	1 0
Ethanol			-.479477	-.507647	-.513854	-154.603229
Boys	15	675	98.200	98.313	98.324	0 0
P-M	8	257	98.207	98.315	98.326	0 0
Hydrazine			-.354552	-.373853	-.377595	-111.555853
Boys	10	210	98.535	98.608	98.622	0 0
P-M	8	127	98.563	98.626	98.740	0 0
Alanine			-.945977	-.973413	-.988156	-322.874523
Boys	54	4770	97.853	98.077	98.026	1 0
P-M	10	1098	98.093	98.247	98.244	1 0
Dioxouracil			-1.186294	-1.197423	-1.218643	-413.697741
Boys	75	6703	98.436	98.599	98.418	3 2
P-M	10	896	98.449	98.603	98.422	0 3

For each molecule, the number of passes and 2×2 rotations required for localization are shown (see text). The percentages of correlation energies recovered by the LC method are shown below the corresponding conventional MP2, MP3, and MP4(SDQ) result. In the last column are shown the total energies ($E^{\text{SCF}} + E^{\text{MP4}}$) as well as the number of three- and four-center LMOs, respectively, as determined by our virtual space selection algorithm (see text). All localizations and correlation calculations were done with frozen core. Our local virtual spaces used only valence shell basis functions. Geometries were optimized using offset forces²⁰ at the 6-31G* level (except for diborane, which is from ref. 10). Energies (atomic units) were calculated at the 6-31G** level.

ceiling of four atoms. This *overall* better representation of systems with delocalized orbitals constitutes a slight advantage of Boys over PML for such systems.

In Table II are presented results that support the assertions of this discussion. Therein are contained the energies from *conventional* Møller-Plesset correlation calculations along with the corresponding percentages recovered when using our *local virtual space* approximation with Boys or P-M LMOs. In all cases, the LC method recovered better than 98% of the energy as calculated by the corresponding conventional method. It should be noted that for furan and benzene (both genuinely delocalized) Boys LMOs are slightly better than P-M LMOs. But, for

all others PML performed better, even for the moderately delocalized dioxouracil.

As for the localizations proper, it is seen that in every case significantly fewer 2×2 rotations were required for PML than for Boys.

All of the local correlation calculations in Table II used LMO virtual spaces of not more than four atomic centers. In the one case of benzene, this restriction meant not meeting the 98% spanning criterion for some of the LMOs, six for the Boys type and three for the P-M. For these four-center virtual spaces, the spanning is 96.4% for Boys and 93% for P-M. For the benzene π orbitals, the natural local space obviously consists of all six carbon centers. However, for larger aromatic systems (naphthalene,

anthracene) delocalization does not significantly increase. For example, in naphthalene there are by PML five delocalized π type orbitals, four of which are spanned by 95.2% with four centers and one of which by 88.5% with four centers. Using Boys localization, there are 10 delocalized τ type orbitals, 8 being 97.6% spanned with 4 centers and 2 93.8% spanned by 4 centers. We expect that a limit of four atomic centers—certainly no more than six—will serve well for most calculations. But, it should be appreciated that the π orbitals in genuinely delocalized aromatic systems may extend over a significant part of the carbon skeleton. Truncation of the virtual space for such orbitals may jeopardize the quality of the calculation somewhat. One may choose, for such cases, to leave the delocalized orbital virtual spaces untruncated. As long as the number of delocalized orbitals is not a large fraction of the total, the advantages of the LC method will still hold.

CONCLUSION

Whether Boys or P-M localization is used, the LC method recovers at least 98% of the correlation energy *as calculated by* a corresponding conventional calculation. For genuinely delocalized systems, the use of Boys localization may be preferred over PML. For physically well-localized systems, PML is superior for use in the LC approach. We expect that for large systems PML will prove to be a significant improvement over Boys when used in this way.

The superiority of PML over Boys LMOs is actually somewhat better than indicated by the percentile figures in Table II. This is because, in general, a large part of the energy ostensibly neglected by the local virtual space approximation is actually not real correlation energy but intramolecular BSSE.^{2-6,11} De-

ducting BSSE from the conventional result causes a widening of the gap between the Boys and P-M percentile figures.

This work was funded by the National Science Foundation under Grant CHE-8814143. The authors also thank IBM Co. for its generous technical support.

References

1. P. Pulay, *Chem. Phys. Lett.*, **100**, 151 (1983).
2. S. Saebo and P. Pulay, *Chem. Phys. Lett.*, **113**, 13 (1985).
3. P. Pulay and S. Saebo, In *Geometrical Derivatives of Energy Surfaces and Molecular Properties*, P. Jorgensen and J. Simons, Eds., Reidel, Dordrecht, 1986, p. 95.
4. P. Pulay and S. Saebo, *Theor. Chim. Acta*, **69**, 357 (1986).
5. S. Saebo and P. Pulay, *J. Chem. Phys.*, **86**, 914 (1987).
6. S. Saebo and P. Pulay, *J. Chem. Phys.*, **88**, 1884 (1988).
7. S. Saebo, *Int. J. Quantum Chem.*, **38**, 641 (1990).
8. S. Saebo, *Int. J. Quantum Chem.*, **42**, 217 (1992).
9. S.F. Boys, In *Quantum Theory of Atoms, Molecules and the Solid State*, P.O. Lowdin, Ed., Academic, New York, 1966, p. 253.
10. J. Pipek and P.G. Mezey, *J. Chem. Phys.*, **90**, 4916 (1989).
11. S. Saebo, W. Tong, and P. Pulay, *J. Chem. Phys.*, submitted.
12. L. Adamowicz and R.J. Bartlett, *J. Chem. Phys.*, **86**, 6314 (1987).
13. L. Adamowicz, R.J. Bartlett, and A.J. Sadlej, *J. Chem. Phys.*, **88**, 5749 (1988).
14. L. Adamowicz, *J. Phys. Chem.*, **93**, 1780 (1989).
15. C. Edmiston and K. Ruedenberg, *Rev. Mod. Phys.*, **35**, 457 (1963).
16. C. Edmiston and K. Ruedenberg, *J. Chem. Phys.*, **43**, S97 (1965).
17. C. Edmiston and K. Ruedenberg, In *Quantum Theory of Atoms, Molecules and the Solid State*, P.O. Lowdin, Ed., Academic, New York, 1966, p. 263.
18. D.A. Kleier, T.A. Halgren, J.H. Hall Jr., and W.N. Lipscomb, *J. Chem. Phys.*, **61**, 3905 (1974).
19. J. Pipek, *Int. J. Quantum Chem.*, **36**, 487 (1989).
20. P. Pulay, G. Fogarasi, X. Zhou, and P.W. Taylor, *Vibrat. Spectrosc.*, **1**, 159 (1990).