Extensions of the S66 Data Set: More Accurate Interaction Energies and Angular-Displaced Nonequilibrium Geometries

Jan Řezáč,[*,†] Kevin E. Riley,[†] and Pavel Hobza[†,‡]

[†]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, 166 10 Prague, Czech Republic

[‡]Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic

Ⓢ *Supporting Information*

ABSTRACT: We present two extensions of the recently published S66 data set [Řezáč, Riley, Hobza; DOI: 10.1021/ct2002946]. Interaction energies for the equilibrium geometry complexes have been recalculated using a triple-ζ basis set for the CCSD(T) term in the CCSD(T)/CBS scheme. This allows for the extrapolation of this term to the complete basis set limit, improving accuracy by almost 1 order of magnitude compared to the scheme previously used for the S66 set. Now, we estimate the largest error in the set to be about 1%. Validation of several methods against the new data indicates the exceptional robustness and accuracy of the SCS-MI-CCSD method. The second extension improves the coverage of nonequilibrium geometries. We introduce a new data set, S66a8, that samples intermolecular angular degrees of freedom in the S66 complexes. For each of the 66 complexes, eight displaced geometries have been constructed, systematically sampling possible rotations of the monomers. Interaction energies in this set are calculated at the CCSD(T)/CBS level consistently with the earlier introduced S66x8 data set that samples the intermolecular distance.

■ INTRODUCTION

The importance of accurate *ab initio* calculations for evaluation of the performance of more approximate methods is now widely recognized. One of the fields where such calculations serve as very important benchmarks is the study of noncovalent interactions. Multiple databases of reference data covering this topic have been published in the past decade.[1−5] Recently, we introduced the S66 data set,[6] which was designed to overcome multiple limitations of the previously available sets. The most important improvements are the increased size of the set and more balanced coverage of different types of interactions, which was achieved by careful selection of the complexes. The interaction energies of these complexes were calculated consistently at the CCSD(T) level and extrapolated to the complete basis set limit (CBS). For a more detailed description of the S66 set, we refer the reader to the original paper.[6]

If these complexes are used as a model for interaction between atomic groups in large molecular systems, it is also necessary to consider nonequilibrium geometries. Although the interaction is strongest in equilibrium, there are much greater numbers of weaker interactions acting over longer distances in large condensed systems. To address this, we have published dissociation curves for all of the 66 complexes of the S66 data set, also calculated at the CCSD(T)/CBS level.[6]

In this work, we present two extensions of the S66 data set. First, we have recalculated the interaction energies in the S66 set using a larger basis set in order to improve the accuracy toward the complete basis set limit. The new benchmark interaction energies are based on extrapolation of the CCSD(T) term from double- and triple-ζ basis sets with diffuse functions on first row

atoms, improving the accuracy over the previously published results by almost 1 order of magnitude. Second, we introduce the S66a8 data set, which samples the intermolecular angular degrees of freedom in all of the 66 complexes. Together with the S66x8 set, sampling intermolecular distances, our results represent the largest body of accurate data available for nonequilibrium structures of molecular complexes, calculated consistently at the same level.

We use the new results as a reference for assessment of the accuracy of more approximate methods. We focus on two methods that have been shown in our previous paper to have very good performance to cost ratios, MP2.5[7] and SCS-MI-CCSD.[8] A comparison of results calculated with different basis sets allows us to discuss the transferability of the parameters used in these methods.

■ METHODS

Basis Sets. Throughout this work, we use the correlation consistent basis sets of Dunning.[9] The full names of the basis sets, cc-pVXZ (X = D,T,Q) and aug-cc-pVXZ (augmented with diffuse functions[10]), are abbreviated as XZ and aXZ in this work.

Benchmark Calculations. The CCSD(T)/CBS interaction energy is obtained as a sum of the Hartree−Fock (HF) interaction energy, the MP2 correlation energy extrapolated to CBS from large basis sets, and a ΔCCSD(T) correction ($\Delta E^{\mathrm{CCSD(T)}} - \Delta E^{\mathrm{MP2}}$) calculated with a smaller basis set. For the large set of
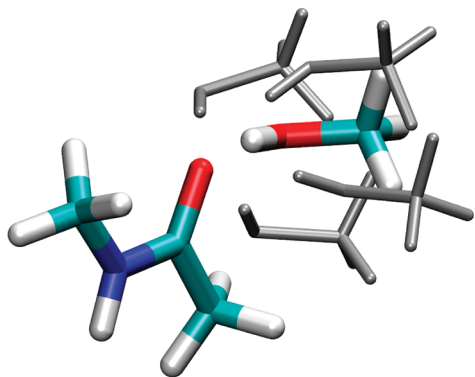
Figure 1. Angular-displaced geometries obtained by rotating one of the monomers (in gray), compared to the original geometry of the complex (in color).

angular displacement calculations, we use the same scheme as for the original S66 and S88x8 sets. The MP2 term is extrapolated from the aTZ and aQZ basis sets, and the ΔCCSD(T) correction is calculated with the aDZ basis set. Extrapolations to the CBS limit are done using the formula of Helgaker et al.[11] All interaction energy calculations presented here are corrected for the basis set superposition error using the counterpoise scheme.

Here, we have recalculated the S66 binding energies using a more accurate scheme, in which the ΔCCSD(T) correction is extrapolated to the CBS from two calculations. Because the calculation of the largest systems in the set using the aTZ basis set is not possible with our current resources, we use a smaller basis set that still retains most of the qualities of the aTZ basis set, combining the TZ basis for hydrogens and aTZ for all other atoms.[12] This basis set is often referred to as the heavy-augmented basis, and we use the abbreviation haTZ. Recently, it has also been published under the name jul-cc-pVTZ.[13] An analogously constructed haDZ basis set is used along with haTZ to extrapolate the ΔCCSD(T) correction. This scheme achieves binding energy values close to the most accurate results available for the S22 set,[14] where aDZ and aTZ basis sets were used for the extrapolation of ΔCCSD(T).

To assess the improvement brought by this approach, we compare these two schemes, along with other possible combinations of basis sets, on a set of small model systems introduced earlier.[15] Description and geometries of these complexes are available in the original paper.

S66a8 Geometries. The angular-displaced geometries have been prepared from the S66 geometries in the following way: For each complex, the principal plane of the monomers was identified. In simple cases, these are defined by symmetry; in more complex molecules, the definition was only approximate but reflects the shape of the molecule. Eight geometries are built for each system: each of the monomers is rotated in both directions ($\pm$) in the molecular plane (coordinate y) and perpendicular to it (coordinate z) by 30°. In order to sample the angular displacements but not the nonequilibrium intermolecular distances (already covered by the S66x8 set), the intermolecular distances in the S66a8 complexes are optimized at the RI-MP2/TZ level, with the counterpoise correction, while all other degrees of freedom are kept fixed. An example of the displaced geometries obtained by the rotation of one monomer is shown in Figure 1. We do not explore rotation around the intermolecular axis, the interaction energy is not sensitive to such a change of the

Table 1. Accuracy of the CCSD(T)/CBS Scheme with Different Basis Sets Used for the Calculation of the ΔCCSD(T) Term, Measured As the Root Mean Square Error in a Set of 10 Small Complexes for Which Accurate Estimates of the CCSD(T)/CBS Interaction Energies Are Available

ΔCCSD(T) basis set(s)	RMSE (kcal/mol)
aDZ	0.080
aTZ	0.020
TZ	0.107
haTZ	0.033
haDZ → haTZ extrapolation	0.009

geometry in most of the studied complexes. No symmetry is assumed in the generation of the displaced geometries; therefore, the complete set contains several pairs of structures built from symmetrical minima that are equivalent in energy. For simplicity, we provide the complete set of geometries, but identical results can be achieved by eliminating the duplicates and setting the weight of the results to two in the statistical analysis.

Methods Tested. There are two methods that we investigate in this work in greater detail; these have been parametrized to describe noncovalent interactions. MP2.5[7] is a variant of MP3 where the third-order contribution is scaled by one-half. SCS-MI-CCSD[8] is a spin-component-scaled CCSD method optimized specifically for noncovalent interactions.

Several other methods have been tested on the S66a8 set. In addition to the ones described above, these are MP2 in multiple basis sets, spin-component-scaled MP2[16] (SCS-MP2) and its reparameterization for noncovalent interactions SCS-MI-MP2,[17] dispersion weighted MP2[18] (DW-MP2), MP3, CCSD, and the original version of spin-component-scaled CCSD[19] (SCS-CCSD). Details on these methods and values of the parameters used can be found in the original paper on the S66 data set.[6]

To extrapolate all of these methods to the CBS limit, we employ a scheme analogous to the CCSD(T)/CBS calculations; the result is built from accurate extrapolation of the MP2 energy and a higher order correction (e.g., $\Delta E^{\text{SCS-MI-CCSD}} - \Delta E^{\text{MP2}}$) calculated in a smaller basis set or extrapolated independently, using the same basis set(s) as the benchmark in the given data set.

Computational Details. Optimization of the intermolecular distance in the S66a8 set was carried out in Turbomole 6.2.[20] All interaction energy calculations were performed using the Molpro 2010 package,[21] using density fitting for the MP2 calculations.

■ RESULTS AND DISCUSSION — S66 DATA SET

CCSD(T)/CBS Extrapolation. In the set of 10 model complexes,[15] we compare the effect of the basis sets and extrapolation schemes applicable to the calculation of the ΔCCSD(T) correction in the S66 set (Table 1). The results are compared to our best estimates of the CCSD(T)/CBS interaction energies directly extrapolated from the aTZ and aQZ basis sets. First, to justify the use of the heavy-augmented basis sets, we discuss the differences between the TZ, aTZ, and haTZ basis sets, also including aDZ for comparison. It is clear that the heavy-augmented basis set yields results close to the fully augmented ones and that the improvement over the TZ basis set with no diffuse functions is substantial. These results are supported by previous studies of this basis set.[12,13]

Table 2. New, More Accurate CCSD(T)/CBS Interaction Energies (in kcal/mol) for the Complexes in the S66 Data Set[a]

	hydrogen bonds	ΔE
1	water···water	−5.01
2	water···MeOH	−5.70
3	water···MeNH₂	−7.04
4	water···peptide	−8.22
5	MeOH···MeOH	−5.85
6	MeOH···MeNH₂	−7.67
7	MeOH···peptide	−8.34
8	MeOH···water	−5.09
9	MeNH2···MeOH	−3.11
10	MeNH₂···MeNH2	−4.22
11	MeNH₂···peptide	−5.48
12	MeNH₂···water	−7.40
13	peptide···MeOH	−6.28
14	Peptide···MeNH₂	−7.56
15	peptide···peptide	−8.72
16	peptide···water	−5.20
17	uracil···uracil (BP)	−17.45
18	water···pyridine	−6.97
19	MeOH···pyridine	−7.51
20	AcOH···AcOH	−19.41
21	AcNH₂···AcNH₂	−16.52
22	AcOH···uracil	−19.78
23	AcNH···uracil	−19.47

	dispersion	ΔE
24	benzene···benzene (π−π)	−2.72
25	pyridine···pyridine (π−π)	−3.80
26	uracil···uracil (π−π	−9.75
27	benzene···pyridine (π−π)	−3.34
28	benzene···uracil (π−π)	−5.59
29	pyridine···uracil (π−π)	−6.70
30	benzene···ethene	−1.36
31	uracil···ethene	−3.33
32	uracil···ethyne	−3.69
33	pyridine···ethene	−1.80
34	pentane···pentane	−3.76
35	neopentane···pentane	−2.60
36	neopentane···neopentane	−1.76
37	cyclopentane···neopentane	−2.40
38	cyclopentane···cyclopentane	−2.99
39	benzene···cyclopentane	−3.51
40	benzene···neopentane	−2.85
41	uracil···pentane	−4.81
42	uracil···cyclopentane	−4.09
43	uracil···neopentane	−3.69
44	ethene···pentane	−1.99
45	ethyne···pentane	−1.72
46	peptide···pentane	−4.26

	others	ΔE
47	benzene···benzene (TS)	−2.83
48	pyridine···pyridine (TS)	−3.51
49	benzene···pyridine (TS)	−3.29

Table 2. Continued

	others	ΔE
50	benzene···ethyne (CH···π)	−2.86
51	ethyne···ethyne (TS)	−1.54
52	benzene···AcOH (OH···π)	−4.73
53	benzene···AcNH₂ (NH···π)	−4.40
54	benzene···water (OH···π.)	−3.29
55	benzene···MeOH (OH···π)	−4.17
56	benzene···MeNH₂ (NH···π)	−3.20
57	benzene···peptide (NH···π)	−5.26
58	pyridine···pyridine (CH···N)	−4.24
59	ethyne···water (CH···O)	−2.93
60	ethyne···AcOH (OH···π)	−4.97
61	pentane···AcOH	−2.91
62	pentane···AcNH₂	−3.53
63	benzene···AcOH	−3.75
64	peptide···ethene	−3.00
65	pyridine···ethyne	−4.10
66	MeNH₂···pyridine	−3.97

[a] The CCSD(T)/CBS is based on extrapolation of the MP2 correlation energy from the aTZ and aQZ basis sets with the ΔCCSD(T) correction extrapolated from the haDZ and haTZ basis sets.

Table 3. Errors of the MP2.5 and SCS-MI-CCSD Methods with Higher Order Terms Calculated in Different Basis Sets Compared to the CCSD(T)/CBS Reference with ΔCCSD(T) Term Calculated in aDZ Basis Set and Extrapolated from haDZ and haTZ Basis Sets

	ΔCCSD(T) basis set in reference	aDZ	haDZ→haTZ
method	basis set	RMSE (kcal/mol)	RMSE (kcal/mol)
MP2.5	aDZ	0.16	0.21
MP2.5	aDZ→aTZ	0.16	0.22
SCS-MI-CCSD	aDZ	0.08	0.14
SCS-MI-CCSD	haDZ→haTZ	0.14	0.07
SCS-MI-CCSD	aDZ→aTZ	0.14	0.07

To improve the accuracy of the final CBS estimate further, it is now possible to extrapolate the ΔCCSD(T) from the haDZ/haTZ series. This scheme yields a root-mean-square error (RMSE) of only 0.009 kcal/mol for the model complexes, lowering the error by almost 1 order of magnitude compared to the scheme used previously (using the aDZ basis set, RMSE 0.080 kcal/mol). The improvement obtained by the extrapolation of the ΔCCSD(T) term, compared to the use of a single calculation in the haTZ basis set, is also substantial. We find this to be the most accurate approach practically applicable even to the largest complexes in the S66 set. Therefore, we consider these results to be the new benchmark for the S66 set.

The largest error in the set of model complexes is 0.7%, compared to 2.5% in the previously used scheme. Therefore, we estimate that the largest errors in the S66 set are lower than 1%.

Benchmark Results. Interaction energies in the S66 data set obtained using the new, more accurate CCSD(T)/CBS extrapolation are listed in Table 2. These results are also available online through the BEGDB database.[22]
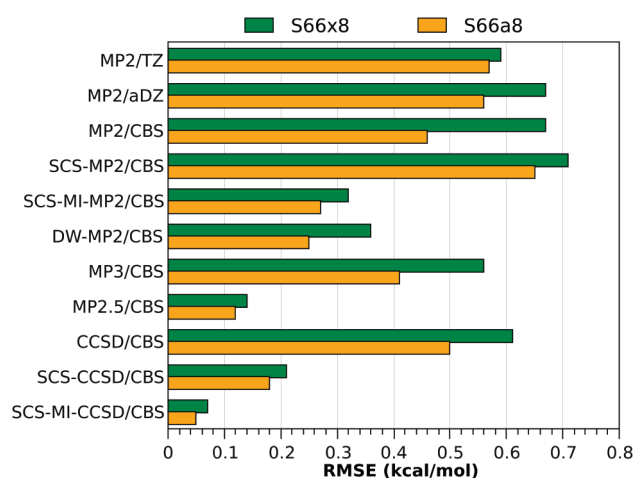
Figure 2. Root mean square error of selected methods in the S66x8 (dissociation curves) and S66a8 (angular displacements) data sets.

Comparing the CCSD(T)/CBS scheme using the aDZ basis set for the ΔCCSD(T) term used previously with the new results, we find an average unsigned error of 0.08 kcal/mol (1.5%) and a RMSE of 0.10 kcal/mol. The strength of hydrogen bonds had been systematically underestimated, while dispersion and mixed-type interactions had been overestimated. This reflects the sign of the ΔCCSD(T) term, whose magnitude is smaller in the unsaturated basis set.

Methods Tested. The overall performance of the methods originally tested on the S66 set does not change significantly when the new, more accurate benchmark is used. Here, we will focus only on the effect of the basis set (used for the higher-order correction in the CBS scheme) on two methods found to be the most accurate in their categories, MP2.5 and SCS-MI-CCSD. Table 3 lists RMS errors in the S66 set for multiple combinations of basis sets, both in the studied method and in the reference.

The performance of MP2.5 is slightly worse when the more accurate reference is used, regardless of the basis set used for the MP2.5 calculation. This indicates that the source of the error comes from the approximations in the method itself and not the basis set. Some improvement can be achieved by optimization of the scaling factor, discussed in detail in a separate paper.[23]

The behavior of SCS-MI-CCSD is surprisingly consistent. This method is able to accurately reproduce the CCSD(T) results calculated using the same basis set or extrapolation scheme (RMSE 0.08 and 0.07 kcal/mol for aDZ and CBS extrapolation). When compared crosswise, the error of the method combined with the difference between the references, the RMSE is larger (0.14 kcal/mol). On the basis of these results, it is obvious that the SCS-MI-CCSD method is very robust and provides results very close to CCSD(T) in a given basis set.

■ RESULTS AND DISCUSSION − S66A8 DATA SET

Benchmark Results. The complete set of S66a8 geometries, benchmark interaction energies, and results from the tested methods are available online in the BEGDB database,[22] where it is possible to browse, plot, and download the data.

The most important information that can be derived from these data is how the magnitude of the interaction in a given complex decreases when the geometry is displaced. We list the average interaction energy per complex as a percentage of the interaction energy in equilibrium for the individual groups of complexes of the S66 set: hydrogen bonds, 77%; dispersion-dominated complexes, 67%; others, 77%. It seems counterintuitive that hydrogen bonds, which are known to be sensitive to the mutual orientation of the interacting molecules, do not exhibit the largest decrease. On the other hand, the hydrogen bonding motif is conserved rather well upon rotation by 30°, while the rotation of larger molecules in the dispersion-dominated complexes leads to a large decrease of the contact surface that determines the strength of the interaction. Dispersion interactions also decay with interatomic distance faster than electrostatic interactions.

Methods Tested. We used the S66a8 data set to test the same set of methods studied on the S66x8 set. The results, plotted in Figure 2 and listed in Table S1 in the Supporting Information, are very similar. Therefore, we refer the reader to the discussion of the performance of individual methods in the original paper.[6] In general, the errors in the S66a8 set are slightly lower even when the relative error (with respect to average interaction energy in the set) is considered. The major source of this discrepancy stems from the shorter than equilibrium geometries in the S66x8 set, where the errors are larger than at (or above) equilibrium distances.

However, the similarity of the results presented here does not make the S66a8 set redundant. It is not surprising that high-level QM methods describe the entire potential energy surface similarly. The main reason for building the S66a8 set was to aid in the development and testing of more approximate methods, such as DFT-D, semiempirical QM methods, or force fields, where significant differences can be expected.

■ CONCLUSIONS

We present new benchmark interaction energies for the S66 data set. Using the extrapolation of the ΔCCSD(T) correction from haDZ and haTZ basis sets, we have improved the average accuracy by almost 1 order of magnitude. On the basis of small model calculations, we estimate the largest error in the S66 set to be approximately 1% of the interaction energy when compared to CCSD(T) complete basis set limit.

The SCS-MI-CCSD method was found to be very robust, as the scaling coefficients are transferable between different basis sets. For a given basis set or extrapolation scheme, the results reproduce reference CCSD(T) calculations using the same basis sets with high accuracy.

Extension of the S66 set to nonequilibrium geometries obtained by rotation of the monomers in the complex, the S66a8 set, is also presented. Here, the high-level QM methods tested yield errors similar to those in the S66x8 data set, but we expect the S66a8 set to be useful for development and validation of more approximate methods where larger differences can be found. When the S66x8 and S66a8 sets are combined, they constitute 1056 points covering the most important coordinates of the intermolecular potential energy surface of the complexes in the S66 set.

Geometries of the complexes, the benchmark CCSD(T)/CBS interaction energies, and results from the other methods discussed in the paper are freely available in the online database at www.begdb.com.[22]

■ ASSOCIATED CONTENT

S **Supporting Information.** Table S1, listing errors of all of the methods tested on S66a8 and S66x8 sets and geometries and

3469

dx.doi.org/10.1021/ct200523a |*J. Chem. Theory Comput.* 2011, 7, 3466–3470

benchmark interaction energies for the data sets featured in this paper are provided. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Fax: +420 220 410 320. E-mail: rezac@uochb.cas.cz.

■ REFERENCES

(1) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(2) Gráfová, L.; Pitoňák, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 2365–2376.

(3) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2011**, *7*, 291–309.

(4) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415–432.

(5) Burns, L. A.; Mayagoitia, A. V.-; Sumpter, B. G.; Sherrill, C. D. *J. Chem. Phys.* **2011**, *134*, 084107.

(6) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.

(7) Pitoňák, M.; Neogrády, P.; Černý, J.; Grimme, S.; Hobza, P. *ChemPhysChem* **2009**, *10*, 282–289.

(8) Pitoňák, M.; Řezáč, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 9611.

(9) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(10) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1994**, *100*, 2975.

(11) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(12) ElSohly, A. M.; Tschumper, G. S. *Int. J. Quantum Chem.* **2009**, *109*, 91–96.

(13) Papajak, E. Zheng, J. Xu, X. Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.,* **2011**; DOI: 10.1021/ct200106a.

(14) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104.

(15) Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 685–689.

(16) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

(17) Distasio, R.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 1073–1083.

(18) Marchetti, O.; Werner, H.-J. *J. Phys. Chem. A* **2009**, *113*, 11580–11585.

(19) Takatani, T.; Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* **2008**, *128*, 124111.

(20) *TURBOMOLE*, v6.2; University of Karlsruhe and Forschungszentrum Karlsruhe GmbH: Karlsruhe, Germany, 2010.

(21) Werner, H.-J. Knowles, P. J. Manby, F. R. Schütz, M.; et al. *MOLPRO*, version 2010.1; Cardiff University: Cardiff, U.K.; Universität Stuttgart: Stuttgart, Germany, 2010.

(22) Řezáč, J.; Jurečka, P.; Riley, K. E.; Černý, J.; Valdes, H.; Pluháčková, K.; Berka, K.; Řezáč, T.; Pitoňák, M.; Vondrášek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261–1270.

(23) Riley, K. E. Řezáč, J.; Hobza, P. Manuscript submitted.