Internuclear separations using least squares and neural networks for 46 new main-group diatomics

Ray Hefferlin

Southern Adventist University, Collegedale, TN 37315, USA

Abstract

Combined least-squares and neural-network forecasts for internuclear separations of main-group diatomic molecules, most with from 9 to 12 atomic valence electrons, are presented. We require that the standard-deviation bounds of the forecasts overlap each other; this requirement is met by 65 molecules, of which 46 seem not to have been studied previously. The composite errors average 0.1036Å on either side of the composite predictions. There is agreement with 33 of 41 independent test data (80.5%), and those not in agreement fall outside the composite error limits by an average of 1.83%.

1. Introduction

It is not the purpose of this work to achieve the precision now achieved by computation of experimental measurement, but to make some reasonably precise and accurate predictions from a periodic system of diatomic molecules. There is evidence for the hypothesis that individual atoms of diatomic molecules echo the periodic law: when the atomic number of either atom passes through an atomic magic number, the data have an extremum; and between successive extrema, main-group data vary similarly [1-3]. This paper reports predictions for spectroscopic constants to single-digit accuracy obtained by using this hypothesis, and thereby also tests the hypothesis. The *presuppositions* are that

1. the optimal independent variables are the period and group numbers of the atoms (R1, C1, R2, C2) – *e.g.*, giving SiO the address (3, 4, 2, 6) – as shown in Fig. 1;
2. portions of the data space are smooth enough to support least-squares and neural-network forecasts of single-digit precision and accuracy (Fig. 1);
3. least-squares and neural-network forecasts with overlapped error measures can be more precise and accurate than either alone.

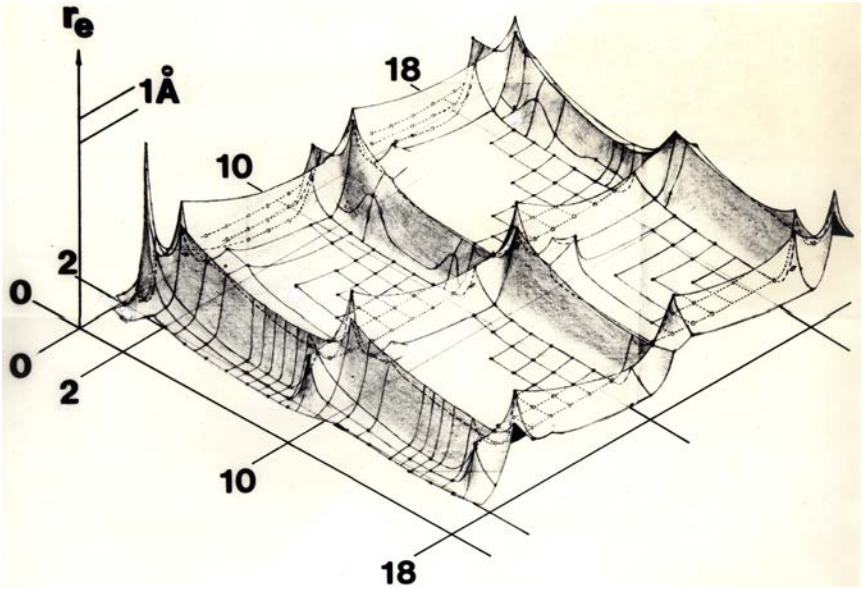

Fig. 1. r_e for diatomic molecules ranging from unipositive hydrides to uninegative period-3 species, plotted on the atomic-number space Z_1,Z_2. A vertical plane on the left-right diagonal passes through homonuclear molecules; the space is symmetric toward and away from the reader with respect to this plane. The ridges pertain to molecules with a rare-gas atom (helium, neon, and argon). The isolated peaks are due to alkaline-earth pairs (Be_2 and Mg_2, and by presumption BeMg and MgBe). Data points are all from [4] except for those taken from the literature to define the flanks of the ridges and the neighborhoods of the small peaks. This behavior continues out to period-6 molecules, so the space (Z_1, Z_2) is cut into subspaces by high ridges bounding square shallow plains. The plains are enumerated by R_1 and R_2, which run from 2 to 6. Within each plain, individual molecules are enumerated by C_1 and C_2; these group numbers run from 1 to 8 because only main-group molecules are under investigation. Reproduced from [3] with permission from Edwin Mellen Press, Lewinston, New York.

The *procedure* consists of constructing a least-squares (LS) formula using internuclear-separation (r_e) data from the 2009 edition of the Handbook of Physics and Chemistry [5]; and training neural network (NN) models on r_e data from Huber and Herzberg's 1979 compilation [4]; forecasting new data, along with error measures, using each method; combining the forecasts and error measures; and testing the combined results with independent literature results. This procedure is similar to that in [6].

Results for 65 molecules whose LS and NN standard deviations overlap are combined, and the average precision is 5.14%, or 0.1036Å when absolute values are averaged. In a test of the accuracy of the predictions, 41 data values, for 22 molecules not in the compilations, were found in the literature; 33 (80.4%) of the data fall within 0.1036Å of the tabulated value and the remainder fall outside by a very few percent.

2. The least-squares analysis

The data in the slightly concave plains of Fig. 1 were fitted with a quadratic polynomial in accordance with presuppositions 1 and 2. For the plain (R_1, R_2) = (2, 2),

$$r_e = 3.53473 - 0.474877C_1 + 0.031992C_1^2 - 0.474877C_2 + 0.031992C_2^2 + 0.0297807C_1C_2 \quad . \qquad (1)$$

This polynomial must be normalized for other values of (R_1, R_2) because the internuclear separations increase with period numbers. Nalewajski and Thakkar [7] show semi-empirically that the normalization should depend on $\log(R_1R_2)$.

The description of procedures in this paragraph closely follows the more detailed description in [6]. Records for main-group molecules were sorted non-redundantly by R_1, R_2, C_1, and C_2. Main-group molecules with atoms from periods 2 to 6 were included, Eq. (1) was used to compute r_e for each set of molecules with fixed period numbers. Good correlations with [4] were obtained for (R_1, R_2) = (2,2) through (2,6), (3,3) through (3,5), (4,4) and (4,5), and (5,5). Within these combinations, when molecules with (C_1, C_2) = (1,1), (2, 2),and (7,7) were excluded the correlations improved markedly. After these selections, the predictions in each (R_1, R_2) were divided by the slope of their trend line, thus producing the *normalized* predictions, hereafter denoted as p. The reciprocals of the original trend-line slopes define the normalization function $f(R_1, R_2)$, Eq. (2) and Fig. 2.

$$r_e = 1.265 - 0.259\ln(R_1R_2) \quad . \qquad (2)$$

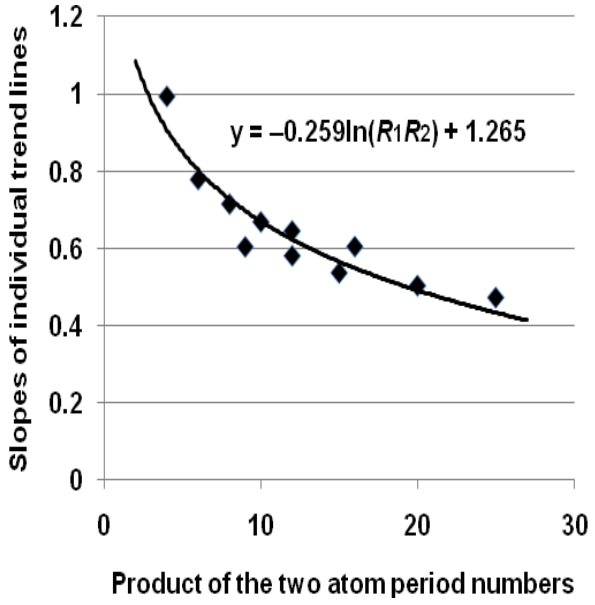

Fig. 2. The behaviors of the slopes of the trend lines of the original r_e predictions, which are used to obtain normalized predictions.

Table 1. Distribution of percent errors, averaged over all R_1 and R_2, the on C_1 and C_2 plane

C_2 \ C_1	1	2	3	4	5	6	7
7	2.576±2.994 15[a]	−0.425±2.155 10	−2.679±2.429 15	−3.707±2.234 7	−1.063±2.657 6	7.994±1.153 4	
6	0.893±5.663 4	4.535±3.287 7	0.064±2.323 4	−1.306±2.380 14	−0.693±2.531 5	2.725±2.750 6	
5			−0.042±0 1	−2.514±2.623 3	−0.480±2.865 4		
4				−4.716±6.299 2			
3			−5.982±0 1				
2							
1							

[a] All entries at lower right of the major diagonal have been combined with, and placed at the locations of, the entries at upper left; there are no entries in column 7, which appears as incomplete.

The normalized data were then resorted according to C_1 and C_2; for each such group-pair the average differences between p and data from [4], and their standard deviations (σ_p), were found and averaged. A separate investigation showed that these measures had no

significant (R_1, R_2) dependence. Table 1 presents the results. The space within the heavy line includes those addresses which have

1. unsigned percent errors of 5% or less
2. σ_p of 7% or less
3. p"σ_p including zero, or a cohort of at least seven data [(2, 6), (3, 7), and (4, 7)]

The data were resorted according to R_1, R_2, C_1, and C_2. They consist of tabulated data from [4]; normalized LS predictions p; σ_p; *adjusted* normalized predictions p' and their standard deviation bounds $p'-\sigma_{p'}$ and $p'+\sigma_{p'}$. The last three quantities are illustrated by considering the group pair (3, 7) in Table 1: the –2.679 indicates that all of the normalized predictions included in this cell should be raised by +2.679% and that the $\sigma_{p'}$ limits then become 0.250% and 5.108%.

3. Inclusion of the neural-network forecasts

The neural-network analysis [8] used data from [4] for training, validation, and testing; several trials (one described in Section 6 of [6]) indicate that the data from [4] and [5] are extremely well correlated. Records for all molecules that are in the final LS results were drawn from the table of NN results which are archived in the supplementary material to [8] (available from the American Chemical Society), and added to the LS file. The additions are the neural-network prediction q and its absolute standard deviation bounds $q-\sigma_q$ and $q+\sigma_q$; the composite (Gaussian average) prediction r_e(*composite*) and its composite error Δr_e(*composite*). The average of this last quantity, which is the random error of the composite prediction process, is 0.136%.

In accordance with the third presupposition stated above, it was determined to keep only those molecular records for which the $\sigma_{p'}$ bounds overlap the σ_q bounds or *vice versa*. For example, the LS bounds for LiSe are 2.062Å and 2.310Å, the NN bounds are 1.912Å and 2.159Å, so there is an overlap. Keeping only these molecules reduces the number of records to 65, with an average Δr_e(*composite*) 0.1036Å. They are listed in Table 2.

Table 2

Internuclear separation values and errors for molecules with overlapped LS and NN standard deviations, sorted alphabetically

Molecule	R1	C1	R2	C2	$r_e(comp)$	$\Delta r_e(comp)$ Absolute	Lower bound	Upper bound
AlSe	3	3	4	6	2.148	0.063	2.085	2.211
AlTe	3	3	5	6	2.286	0.078	2.208	2.364
AsI	4	5	5	7	2.510	0.067	2.443	2.577
AsSb	4	5	5	5	2.262	0.065	2.197	2.326
AsSn	4	5	5	4	2.413	0.089	2.324	2.503
AsTe	4	5	5	6	2.353	0.091	2.262	2.444
BeAt	2	2	6	7	2.187	0.089	2.097	2.276
BeI	2	2	5	7	2.101a	0.077	2.024	2.178
BeSe	2	2	4	6	1.858	0.081	1.777	1.940
BeTe	2	2	5	6	2.017	0.116	1.901	2.133
BPo	2	3	6	6	1.909	0.056	1.853	1.965
BrSb	4	7	5	5	2.510	0.067	2.443	2.577
CAs	2	4	4	5	1.661	0.062	1.600	1.723
CaTe	4	2	5	6	2.637	0.110	2.526	2.747
CBi	2	4	6	5	1.854	0.056	1.798	1.910
CGe	2	4	4	4	1.742	0.177	1.566	1.919
ClSb	3	7	5	5	2.354	0.063	2.291	2.416
ClSr	3	7	5	2	2.594	0.071	2.523	2.665
CPb	2	4	6	4	1.958	0.169	1.789	2.127
CSn	2	4	5	4	1.853	0.199	1.654	2.052
CTe	2	4	5	6	1.760	0.051	1.708	1.811
FBi	2	7	6	5	1.957	0.052	1.905	2.009
GaTe	4	3	5	6	2.442	0.078	2.364	2.520
GeSb	4	4	5	5	2.414	0.089	2.324	2.503
GeSn	4	4	5	4	2.523	0.204	2.319	2.728
KTe	4	1	5	6	2.947	0.336	2.610	3.283
LiSe	2	1	4	6	2.111	0.199	1.912	2.310
LiTe	2	1	5	6	2.088	0.007	2.081	2.095
MgTe	3	2	5	6	2.501	0.132	2.369	2.633
NaS	3	1	3	6	2.588	0.147	2.442	2.735
NaSe	3	1	4	6	2.649	0.196	2.453	2.845
NAt	2	5	6	7	1.994	0.089	1.905	2.083
NBi	2	5	6	5	1.950	0.051	1.900	2.001
NGe	2	5	4	4	1.673	0.050	1.623	1.723
NI	2	5	5	7	1.863	0.075	1.788	1.939
NPb	2	5	6	4	1.884	0.071	1.812	1.955
NPo	2	5	6	6	1.855	0.091	1.763	1.946

NSb	2	5	5	5	1.702	0.049	1.653	1.750
NSn	2	5	5	4	1.783	0.060	1.724	1.843
NTe	2	5	5	6	1.721	0.069	1.652	1.790
OTl	2	6	6	3	1.893	0.072	1.821	1.965
PCl	3	5	3	7	2.091c	0.056	2.035	2.146
PS	3	5	3	6	1.933	0.049	1.884	1.982
PSn	3	5	5	4	2.254	0.074	2.180	2.328
SAs	3	6	4	5	2.059	0.100	1.960	2.159
SbI	5	5	5	7	2.675	0.071	2.604	2.746
SbTe	5	5	5	6	2.460	0.091	2.369	2.551
SeIn	4	6	5	3	2.432	0.089	2.343	2.520
SeRb	4	6	5	1	2.942	0.067	2.876	3.009
SeSb	4	6	5	5	2.361	0.099	2.262	2.459
SeSr	4	6	5	2	2.613	0.086	2.527	2.699
SeTe	4	6	5	6	2.355	0.091	2.264	2.446
SGa	3	6	4	3	2.144	0.059	2.085	2.203
SiGe	3	4	4	4	2.233a	0.140	2.083	2.363
SIn	3	6	5	3	2.272	0.092	2.180	2.364
SiP	3	4	3	5	1.992	0.071	1.920	2.063
SiSb	3	4	5	5	2.253	0.073	2.180	2.327
SiSi	3	4	3	4	2.082a	0.191	1.891	2.272
SiSn	3	4	5	4	2.358	0.200	2.157	2.558
SiTe	3	4	5	6	2.235	0.083	2.152	2.318
SK	3	6	4	1	2.637	0.208	2.429	2.845
SnSb	5	4	5	5	2.520	0.091	2.429	2.611
SnSn	5	4	5	4	2.623a	0.284	2.339	2.907
SRb	3	6	5	1	2.633	0.305	2.328	2.939
SSb	3	6	5	5	2.210	0.089	2.121	2.299

(a) Non-standard configuration [9]

4. Accuracy

Three tests of the composite predictions were performed to determine if the composite predictions are accurate:

a) Of the 65 composite predictions, there are data for 14 of them in [9]; of these, 10 data agree and four do not [they lie outside of Δr_e(*composite*) by 5.27%, 2.33%, 0.09%, and -1.64%] respectively.

b) Of 24 data for six molecules given by [10], 21 agree and three do not [all three relate to one molecule and are 1.47%, 1.15%, and 0.74% outside of Δr_e(*composite*)]. One of the 21 in agreement has an anomalous *aufbau* [9].

c) Of three data from the new section of high-precision values in the Web site [4], two agree and one does not [2.17% outside of Δr_e(*composite*)].

In sum, there is agreement with 33 of 41 test data (80.5%), which is consistent with the conjecture that Δr_e(*composite*) exceeds one "standard deviation." The agreements might have been even a larger percent of the 41 had the errors associated with the literature data been included.

5. Discussion

There are 46 composite predictions for which no tabulated or literature data were found. Do these contribute truly new information? The answer is not known because data for some of these molecules might not have been found in our searches of the literature. However, the last section suggests that there is an 80.5% chance that any new datum will agree with the combined predictions given here to within Δr_e(*composite*) — and that if it does not, then it will fall outside Δr_e(*composite*) by a very few percent.

There are molecules with anomalous electron configurations; do they affect the results? Test (a) above: of the 10 molecules with composite predictions in agreement, two have anomalous *aufbau*; of the four in disagreement, one (0.09% outside) has anomalous *aufbau* [9]. Test (c): one of those in agreement has an anomalous *aufbau* [9]. It is concluded that having an anomalous configuration has not affected the LS and NN prediction process significantly.

Current experimental or computational predictions for r_e are far more precise than those presented here and in some decades will replace them. In the meantime, it is hoped that the data given here might be useful as first approximations; that the joint use of LS and NN, common in mathematical chemistry, will have been shown useful in another area; and that the original hypothesis, as manifested in periodic systems of molecules using period and group axes [3, 11, 12], will have been further supported.

Acknowledgements

The author thanks Professor A. I. Boldyrev (Utah State University) for an alert concerning non-*aufbau* electronic configurations; undergraduate student Ms. Amy Beard for her assistance in the work through part of Section 2; a reviewer for his justifiable criticisms, and the Southern Adventist University Endowed Fund for International Research in Physics.

References

[1] Shchukarev S. A., Neorganicheskaya khimiya, Vol. 1, Vysshaya Shkola (Moscow), 1970.

[2] Kong F.-A., The Periodicity of Diatomic Molecules, *J Mol Struct*., 1982, 90, 17-28.

[3] Hefferlin R., Periodic systems of molecules and their relation to the systematic analysis of molecular data, Edwin Mellen Press, Lewiston, New York, 1989.

[4] Huber K. P., Herzberg G., Constants of Diatomic Molecules, Van Nostrand Reinhold, 1979. Data for individual molecules are available on-line, courtesy of the National Institute of Standards and Technology, by going to the site:

http://www.webbook.nist.gov/chemistry.

[5] Lide D. R., Editor in chief, Handbook of physics and chemistry. Taylor and Francis;2009.

[6] Hefferln R., Vibration frequencies using least squares and neural networks for 50 new *s* and *p* electron diatomics, *J. Quant. Spectr. Radiat. Transf.*, 2010, 111, 71-77.

[7] Nalewajski R. F., Thakkar A. J., Correlations between average atomic numbers and spectroscopic constants of diatomic molecules, *J. Phys. Chem.*, 1983, 87, 5361-5367.

[8] Hefferlin R., Davis W. B., An atlas of forecasted molecular data. 1. Internuclear separations of main-group and transition-metal neutral gas-phase diatomic molecules in the ground state, *J. Chem. Inf. Mol. Model.*, 2006, 46, 820-825.

[9] Boldyrev A. I., Simons J., Periodic table of diatomic molecules, wall chart A, Wiley, 1997.

[10] Ruette F., Sanches M., Añez R., Bermúdez A., Sierraalta A., Diatomic molecule data for parametric methods, *Int. J. Mol. Struct. THEOCHEM*, 2005, 729, 19-37.

[11] Hefferlin R., Zhuvikin G., Caviness K., Duerksen P. J., Periodic systems of *N*-atom Molecules, *J. Quant. Spectr. Radiat. Transf.*, 1984, 32, 257-268.

[12] Hefferlin R., Matrix-product periodic systems of molecules. JQRST 1994, 34, 314-317.