An Atlas of Forecasted Molecular Data. 1. Internuclear Separations of Main-Group and Transition-Metal Neutral Gas-Phase Diatomic Molecules in the Ground State

Ray Hefferlin,*,† W. Bradford Davis,‡ and Jason Ileto†

Southern Adventist University, Collegedale, Tennessee 37315, and Davis Research Group, Rio Linda, California 95673

Received July 7, 2002

Needed spectroscopic data on diatomic molecules can often be found in the superb critical tables of Huber and Herzberg or in the literature published since 1979. Unfortunately, these sources apply to only a fraction of the diatomic species that can exist and so investigators have had to rely on interpolation, additivity, or ad hoc rules to estimate needed values, all of which require other information that is often lacking. This Atlas presents 1001 additional internuclear separations for use until critical tables are available to fill the needs more precisely. The Atlas was produced by mining the data from Huber and Herzberg for trends with least-squares analysis and with neural network software. There are 162 molecules about whose data Huber and Herzberg had no qualifications and whose data were employed for this work; 248 copies of data with low and high magnitudes were added to reduce the effects of frequency. Internuclear separations for 1001 species not found in Huber and Herzberg are presented, and least-squares predictions supplement some of them. The results, i.e., the Atlas, are presented as Table A, Supporting Information. The average error, based on the average of the absolute differences between the predicted values and tabulated values for the molecules having Huber and Herzberg data, is 0.074 Å; if each error is expressed as a percent of the forecast to which it pertains, the average of these errors is 2.94%. There are 25 "questionable" data from Huber and Herzberg, not used in the preparation of the Atlas, for which predictions are included in the Atlas. Of these, 14 agree with the predicted internuclear separations to within twice the stated errors. Additional atlases for other properties of diatomic molecules are in preparation.

1. INTRODUCTION

1.1. The Need for Globally Predicted Data and the Role of This Atlas. Data for diatomic molecules are needed for many purposes. These include such esoteric uses as the astrophysics of stellar atmospheres and interstellar space and the spectroscopic diagnostics of various kinds of flames and arcs. The data may also aid in the understanding of reactions relating to emissions into the atmosphere, sea, and earth and relating to the design of cleaner, more efficient, and hopefully "renewable" fuels for transportation and power generation.

Huber and Herzberg[1] provided a splendid compilation of critically analyzed spectroscopic constants in 1979, and new data are continually made available to the scientific community via the literature. For example, the *Journal of Physical Chemistry* alone publishes on average a dozen articles presenting additional internuclear separations for molecules not included in Huber and Herzberg every year. This estimate is based on entries in 150 issues of the bimonthly "Berkeley Newsletter,"[2] which cites articles having anything to do with small gas-phase molecules. There are many other journals, and so the total made available might be a hundred new internuclear separations each year. Even so, if all these new data were as dependable as those in Huber and Herzberg the progress would be very slow, given that the total number of diatomic molecules formed of atoms with $1 \leq Z \leq 118$ is 7021. Individual researchers who fail to find needed internuclear separations from these sources must wait until experiments or computations are done, or interpolate values using quality data for nearby molecules (if available), or apply rules such as that of Walsh (if quality data for the other property of the same molecule are available), or resort to additivity (if the two atoms are among those for which additivity has been, in principle, established).

The data in this Atlas of globally predicted internuclear separations for neutral diatomic molecules (Table A, Supporting Information) are meant to fill some of the huge void just described. It should have practical use in the endeavors described above even though the predictions contained in it are of lower precision than that which can be achieved by experiment or computation. The Atlas culminates preliminary work on forecasting molecular properties using least-squares methods[3,4] and neural networks.[4,5]

1.2. Preceding Contribution. Neural-networks were first successfully trained and tested on molecular data at the University of Memphis (UM);[6] no global predictions were made. Data for 199 diatomic internuclear separations (r_e) were included in the beautiful study. The data pertain to molecules formed from main-group elements and include molecules containing the two heavy isotopes of hydrogen, one alkaline-earth dimer (Mg_2), and four molecules containing a row-7 atom (YbF, LaO, LaS, and LuD). 19 of the

* Corresponding author phone: (423)238-2869; fax: (423)238-2349; e-mail: hefferln@southern.edu.
† Southern Adventist University.
‡ Davis Research Group.

10.1021/ci020291q CCC: $25.00
Published on Web 02/01/2003

species are positive ions. Molecules with rare gas atoms and with split ground states were excluded. There are six independent variables (inputs) for each of the two atoms: the atomic weight in amu, the atomic number, and the numbers of valence s, p, d, and f electrons. The sum of the atomic charge numbers is the 13th input (independent variable).

QUICKPROP, an upgrade of which is known as BRAINCEL (Jurik Research), was used for the UM investigation. Some 2000 models were trained with 5000 trials each. This procedure likely resulted in "memorization" of the data. Neural nets that have memorized the input data do not generalize well and are thus unable to make good predictions.

2. THEORY

The basis set for this work chosen without knowledge of the UM work. It consists of part of the coordinates of the periodic system of diatomic molecules.[7,8] These coordinates consist of the period number, or row number (R), and group number, or column number (C), of each atom's location in the chart of the elements. These four independent variables were used for the least-squares portion of the computations. The errors resulting from using functions of these four variables are far less than the errors associated with correlations of molecular properties to functions of what would seem a more obvious choice, the atomic numbers.[3,9] The square and cube of each variable are added to form the 12 inputs for the neural-network portion of the work, for the reason to be stated in Section 3.1.2.

Diatomic molecules whose atoms come from different rows of the periodic chart but have the same valences are "isovalent" to the original molecules. Nalewadjski and Thakkar have shown that r_e of isovalent molecules are reasonably well correlated with atomic numbers by the theory-based equation[10]

$$r_e = A + B \log (R_1 R_2) \qquad (1)$$

where A and B depend on the column numbers, in agreement with exhaustive graphical and curve-fitting studies.[11,12] This dependence reflects the behavior of the data as shown, for instance, in Figure 1.

Molecules with main-group atoms from different groups in the same row(s) appear, upon graphical inspection, to form slightly elliptical convex valleys pierced and bounded by high features due to van der Waals species; these flat valleys can be characterized by polynomials of up to the second power.[11,12] Such equations are of use in a least-squares approach but are not asked for in a neural-network approach.

The neural networks are given the inputs (R_1, C_1, R_2, C_2, and the squares and cubes of each of these) and the tabulated data. The inputs can be reversed without changing the identity of the molecule. For this reason, each original tabulated datum for the heteronuclear molecular formula AB was duplicated in reverse for the form BA, i.e., the data were "symmetrized." To check whether an ensemble of data has been properly symmetrized, row and column centroids are used. They are defined as follows

$$\langle R_1 - R_2 \rangle = \sum [r_{ei} \times (R_{1i} - R_{2i})] / \sum r_{ei} \qquad (2)$$

where r_{ei} are tabulated or predicted data and R_{ni} are the period numbers of the atoms in molecule i, and

$$\langle C_1 - C_2 \rangle = \sum [r_{ei} \times (C_{1i} - C_{2i})] / \sum r_{ei} \qquad (3)$$

where C_{ni} are the group numbers of the atoms in molecule i. The centroids for the original data set are zero, to within rounding errors (-2.10×10^{-17} and -1.10×10^{-16}, respectively).

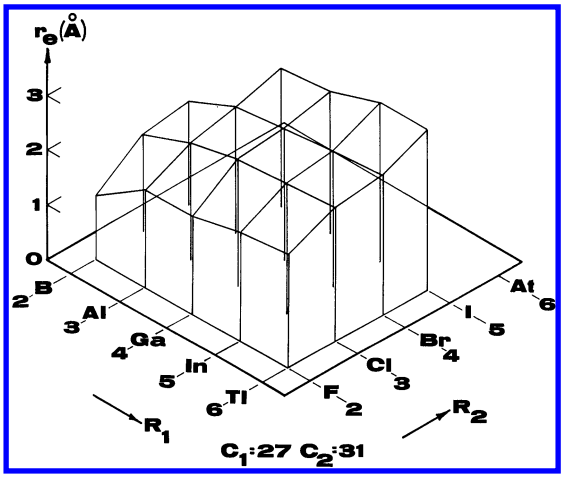

Figure 1. Tabulated internuclear separations of a sample set of isovalent, neutral, ground-state, diatomic molecules. The terrain is such that an estimate of r_e for BI can be made with dividers, given that there is no perspective in the graph. The slight decreases of slope for $R_i = 3$ and 5 are related to period-doubling in the periodic chart of the elements. There is no term for this small effect in eq 1.

3. RESULTS FOR r_e

3.1. The Neural-Network Study. 3.1.1. Data and Inputs. The internuclear separations included in this study pertain to gas-phase, neutral, ground-state diatomic molecules. Only Huber and Herzberg data[1] for r_e were used for this investigation, thus avoiding major uncertainties that often exist in using data when they are first published. The inputs are the row numbers R_1 and R_2 (with domains from 2 to 6) and the column numbers C_1 and C_2 (from 1 to 18, but in practice from 1 to 17 because rare-gas atoms are not included). Rare-gas molecules and molecules with $(C_1,C_2) = (2,2)$, alkaline-earth pairs, were excluded because their internuclear separations are usually larger than those of their neighbors in the C_1,C_2 plane and hence complicate the training of the network. Also excluded were internuclear separations for the molecules denoted in Huber and Herzberg as less certain by the use of parentheses, brackets, or both. Also, Bi_2, CCl, CuO, and TiS were simply overlooked in data entry. 162 molecules remain; 20 of them are homonuclear molecules; 77 molecules contain only main-group atoms; and 84 contain one or two transition-metal atoms. Isotopes, including deuterium and tritium, were ignored. All heteronuclear-molecular data were entered twice (for the reason to be stated at the end of this section), once in AB order and once in BA order, making 304 entries. The average value of r_e is 2.12 Å.

When these entries were sorted in order of increasing magnitude, it became clear that the number of data per unit magnitude is approximately constant in the middle of the domain and much smaller at the low and especially at the

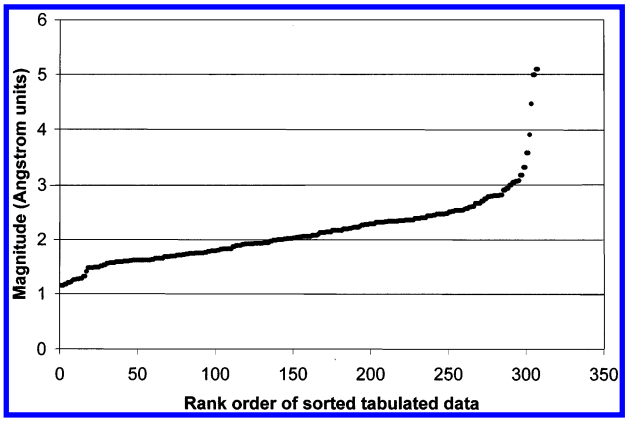

Figure 2. Tabulated data for internuclear separations, before frequency compensation, sorted in order of increasing magnitude.

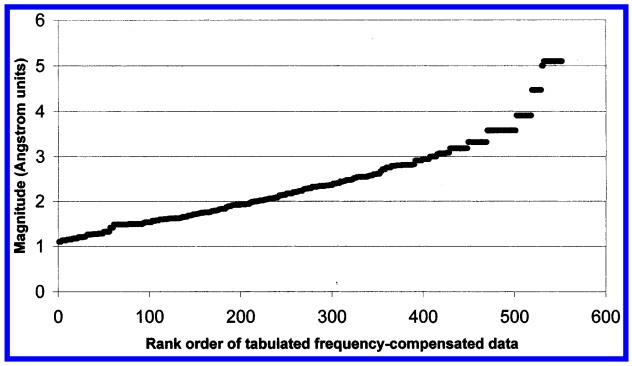

Figure 3. Tabulated data for internuclear separations, after frequency compensation, sorted in order of increasing magnitude.

higher end (Figure 2). These deficits would cause the low and high value predictions made by the program to be susceptible to large errors. To reduce this susceptibility, 248 duplicate values of the entries (inputs and tabulated data) were introduced at the low and high ends of the domain to form a larger set of inputs and data (Table 1). This process is referred to as "frequency compensation". There are, as a result, 552 entries. The average of r_e for these frequency-compensated entries, 2.65 Å, is higher than that for the 304 original entries because many more were duplicated at high values of the internuclear separation than at low values.

There are two problems with frequency compensation. One problem is the creation of plateaus by the act of introducing the duplicate information. The second is that symmetry is difficult to maintain during the frequency compensation. It is often necessary to add an odd number of copies of tabulated data (and their inputs) with a given magnitude because one or both centroids would increase more than expected upon the insertion of the one more copy needed to match an added formula AB with an added formula BA. The end result is a distribution of frequency-compensated data per unit magnitude in which the very steep portions have been flattened as much as possible without the introduction of overly wide plateaus (Figure 3) or unnecessary bloating of the centroids for the learning set.

3.1.2. The Computer Program. Neuralwork's PREDICT was used for neural-network model building in this investigation.[13] The 552 entries, after frequency compensation, were partitioned by the investigators into a "learning set" (512 entries) and a "validation set" (40 entries); care was taken that molecules from grossly underrepresented portions of the input data space were not used in the validation set.

Table 1. Values Added for Frequency Compensation

magnitude	tabulated values	added values
1.1	1	2
1.13	2	4
1.15	2	4
1.17	2	4
1.2	2	4
1.21	1	2
1.24	1	
1.26	2	4
1.27	2	4
1.28	2	4
1.32	2	5
1.41	1	3
1.48	4	12
1.49	4	12
1.51	2	
1.53	2	6
1.56	2	2
1.57 to 1.92	91	
1.93	6	2
1.94 to 2.52	121	
2.54	6	6
2.56	1	1
2.57	2	2
2.6	2	2
2.61	2	2
2.67	2	
2.71	2	2
2.75	2	4
2.79	2	4
2.8	2	4
2.81	4	8
2.82	2	
2.91	2	6
2.94	2	6
3	2	6
3.05	2	
3.07	2	6
3.08	1	3
3.18	2	18
3.32	2	19
3.58	2	30
3.91	1	16
4.47	1	10
5	2	
5.1	2	19
subtotal	304	248
total		552

Using such a large percentage for training is not unusual when there are so few data compared to normal neural-network applications.

The program selects 80% of the training set with which to build a model and uses the rest to test the results. At regular (unstated) intervals, it changes which molecules are in the 80% set and which are in the test set. Thus, the construction of a model involves the entire training set, even though at each interval between "key points" (not specified in the documentation) in time 20% of the training-set data are used for testing. The test consists of comparing the overall correlation between the input and the predicted data of the model being built against the overall correlation of the previous model at the key points, and the result is used to determine if adding more than one node at a time is a better strategy than adding one node at a time.

The neural network program was allowed to create its own independent variables based on common linear and nonlinear transforms of the inputs and to choose from the resulting inputs the "best" set. To encourage its use of at least some

measure of each of the four independent variables R_1, C_1, R_2, and C_2, their squares and cubes were included, making a possible total of 12 user-defined independent variables. Numerous trials showed that this addition results in much more symmetric errors.

The program adds one, two, or three nodes at a time as determined by the strategy. At each addition step, it selects one from a list of transfer functions (sigmoid, Gaussian, sine, and tanh, the last of which is on the list three times because it is the most likely one to work well). The patience was set to six, i.e., the program stops its addition of one or more nodes after six tries if there is no improvement greater than a required measure (0.00001) in the overall correlation. In other words, an improvement of 0.0001 or more (the tolerance) is sufficient to prevent cessation of computation after six tries. This number six (the patience) is chosen so that all transfer functions are attempted and tanh is attempted three times as often. Even if there is improvement exceeding the tolerance, it stops model building after 18 tries. Backpropagation is controlled on the basis of standard correlation error measure. A stochastic factor is added at each try, to prevent the model's settling into a local minimum correlation. This noise factor was varied, but a common value was 0.15.

A model, then, consists of a number of nodes having transfer functions that are three times as likely to be tanh as any of the others. The program goes on to create at most 18 models, but it will stop trying to produce a new model after failure to achieve an improvement of 0.00001 (also called the tolerance) after six consecutive attempts (this six is also called the patience).

The best model of the run is chosen by an internal algorithm that uses a combination of the best overall correlation and the least number of nodes. The standard correlation margin by which the best model has that status is often very small (a couple of thousandths). The smaller the number of nodes, the more likely the model will be to generalize to a larger number of predictions.

The best model of all the runs then predicts values for the molecules in the validation set. Finally, the best model is asked to make the global predictions.

3.1.3. Analysis of the Model. The training of the best model (defined above) is considered successful because the average of the absolute percent errors for the validation set is only a factor of 1.500 different than that of the training set and is in fact smaller (Table 2). The investigators also consider it a success given that (a) the sensitivities and their variances for any pair of conjugate inputs (e.g., R_1 and R_2 or C_1^2 and C_2^2) are not dissimilar by more than a factor of 5, that (b) at least one power of C_1 (which is C_1^2) and one power of C_2 (C_2) is close (17%) to any power of R_1 (R_1^2) and to any power of R_2 (R_2), respectively, indicating that they were all used by the model, and that (c) the averages of the predicted data for the validation and training sets are quite similar (2.415 and 2.460 Å). The sensitivities are the numerical partial differentials (with much smaller than unit denominators) of predicted r_e in the directions defined by the 12 independent variables.

It is also noteworthy that there are only six (really five) molecules with differences between their predicted and tabulated values exceeding 20%; those in the training set are CsHg, MgAu, and RhC and those in the validation set are AuBe, BeAu, and PtO. Except for AuBe, the reversed molecules had predicted data that differ from the tabulated data by less than 20%. Table 2 gives additional statistical information about the training and validation sets for the model. The average of the unsigned errors is greater than the median for the training set but is less than the median for the validation set. These relations indicate that there must be relatively more small-error data in the validation set than in the training set.

Table 2. Statistical Analysis of the Frequency-Compensated Neural-Network Model

Learning Set	
number of data[a]	512
minimum R	2
maximum R	6
minimum C	1
maximum C	17
minimum tabulated datum	1.1
maximum tabulated datum	5.1
minimum prediction	1.230
maximum prediction	5.060
average % difference from tabulated data	−0.380
standard deviation	5.001
average | % difference from tabulated data |	3.623
standard deviation	3.464
median	2.607
Validation Set	
number of data[a]	40
average % difference from tabulated data	−3.992
standard deviation	18.697
average | % difference from tabulated data |	2.415
standard deviation	16.818
median	4.408
Global Predictions	
number of predictions[a]	4220
minimum R	2
maximum R	6
minimum C	1
maximum C	17
minimum prediction	1.230
maximum prediction	5.060
average prediction	2.815
standard deviation	0.900
median	2.564

[a] Counting heteronuclear molecules twice (AB and BA forms).

The predictions for the 552 frequency-compensated entries were ranked in order of increasing magnitude. Bins were defined with lower bounds of 1.1, 1.3, 1.5, ... Å; the average of the differences from the tabulated data were established for each bin (Table 3). There are four bins with one datum or no data; the average of the differences for these bins was calculated as the weighted average of the averages for the other bins. These averages are later applied to the globally predicted data, for which no better error measure exists.

3.1.4. Global Predictions. The entries in Table A, Supporting Information, originated from 7225 [(17 groups) × (5 periods) squared] original global predictions. 25 alkali-earth pairs, (C1,C2) = (2,2), were then culled out, leaving 7200 entries, including heteronuclear molecules twice (with forms AB and BA). 3000 molecules that do not exist (C_1 or C_2 from 3 to 12 in R_1 and R_2 equal to 2 and 3) were next culled, leaving 4200 entries. Additional entries were later culled for the reasons given in Section 4.

The molecular formulas for homonuclear molecules contain both atomic names even when they are the same as in AlAl. Aside from this notation, the formulas appear alpha-

Table 3. Average Absolute Differences between Predictions and Tabulated Values for 552 Predictions, Enhanced in Cases of Small Numbers of Data, Arranged in Bins of Increasing Value

lower limit of predicted data bin (Å)	number of data	average absolute difference between predictions and tabulated data (Å)[a]
1.1	51	0.068
1.3	53	0.033
1.5	50	0.061
1.7	46	0.072
1.9	37	0.101
2.1	25	0.096
2.3	46	0.129
2.5	42	0.114
2.7	37	0.082
2.9	32	0.082
3.1	35	0.062
3.3	30	0.144
3.5	16	0.057
3.7	18	0.254
3.9	0	0.096
4.1	1	0.096
4.3	0	0.096
4.5	11	0.284
4.7	0	0.096
4.9	22	0.167

[a] All bins with no data or one datum have 0.096 Å, i.e., the weighted average of the other values.

betically just as in Huber and Herzberg, with the result that a series of molecules having the same first atom can sometimes be interrupted by molecules having a different first atom (e.g., BaRh, BAs, BaSe). The predictions for the two forms AB and BA are shown in columns 6 and 8. That they differ is no surprise given that a neural-network model is used, and the phenomenon results in the remarkably small *R*- and *C*-centroids of 0.00592 and −0.0109, respectively. With each prediction is its error, as given in Table 3. The average of r_e for globally predicted entries is 2.94 Å, reflecting the presence of more numerous predictions for heavy molecules, relative to the distribution of the tabulated data (even after frequency compensation). The minimum and maximum predictions for all of the prediction sets are within or equal to the lower and upper limits, respectively, of the tabulated data (Table 2). The remaining contents of Table A, Supporting Information, and an explanation of how the individual predictions and their error bars are combined into the final results are discussed in Section 3.2.2.

3.2. The Least-Squares Study. 3.2.1. Smoothing the Tabulated Data. The same set of molecules was included as for the neural-network study except that molecules with transition-metal atoms were not included. The inputs consisted of the row numbers R_1 and R_2 (defined as before) and the group numbers C_3 and C_4 (enumerated from 1 to 7, in contrast to C_1 and C_2 which were enumerated from 1 to 17). This change of numeration took place because only main-group molecules were included: fitting equations to their data precisely was hard enough and fitting equations precisely to the data for transition-metal molecules proved impossible. The set was not symmetrized.

Determination of the fitting equation(s) was done in two parallel steps. For fixed-column molecules, the equation used is

$$r_e = K_0 + K_1(\log R_1) + K_2(\log R_2) \qquad (4)$$

clearly a generalization of eq 1. The number of molecules for each pair of row numbers is not large; in fact only for $(R_1,R_2) = (2,2)$, (2,3), and (3,3) were there enough tabulated data to determine the coefficients well. For fixed-row molecules, the tabulated data were tested against many equations to determine a best fit for all (C_1,C_2).[3] The surviving expression is

$$r_e = K_3 + K_4C_3 + K_5C_3^2 + K_6C_4 + K_7C_4^2 + K_8C_3C_4 \qquad (5)$$

For several pairs of column numbers, the numbers of molecules were too small to determine the coefficients well.

The coefficients of both equations are given in Table 4. The third column shows how many molecules had tabulated data that could be used in the smoothing. The differences between the predictions and the tabulated data, for all of the entries, were ranked in order of increasing magnitude. Bins were defined, and the average of the differences between the forecasted and the tabulated data was established for the entries in each bin; these averages were all very close, and their average is 0.135 Å.

3.2.2. Global Predictions. Fixed-row and fixed column predictions, for molecules not in the Huber and Herzberg compilation, appear in columns 10 and 11 of Table A, Supporting Information. Two entries had both fixed-row and fixed-group predictions; they are for the same molecule (PCl and its reflection ClP, which therefore has its r_e predicted four times). There were 12 molecules with fixed-row and 27 molecules with fixed-column predictions, making a total, after adjustment for PCl, of 38 molecules. All of these species had neural-network forecasts. The same least-squares values exist in both the AB and the BA forms of the molecular symbols.

4. THE ATLAS OF GLOBAL PREDICTIONS FOR r_e

There were 4220 entries left after the culling described Section 3.1.4. Of these, 2255 were eliminated because the values for the AB or BA forms differed each other by more than the sums of their errors (eliminating the possible memorization effects); or because values for the AB or the BA forms differed from the fixed-row or the fixed-column least-squares estimates by an amount exceeding the sums of their errors; or because the fixed-row and fixed-column least-squares forecasts differed by an amount greater than the sum of their errors; or because they had Huber and Herzberg data used for the mining process. Of the 1965 entries that remain, 37 pertain to homonuclear molecules and 1928 pertain to 964 heteronuclear molecules; thus, in total, 1001 molecules are represented. As described in Section 3.1.4, Table A, Supporting Information presents the final results with the molecules in an alphabetical order. Each heteronuclear molecule appears twice, with identical information except that the atomic symbols and the neural-network predictions are reversed. This arrangement makes the table easier to use—it does not matter which of the two appearances is selected. The first five columns in the table give the names and coordinates of the molecules. The next four columns present the neural-net work results and their errors as described above. The next columns give the least-squares results for fixed-row molecules and for fixed-column molecules; the error limits for the entries in these columns are all 0.135 Å.

Table 4. Coefficients for Eqs 4 and 5

molecules	examples	data	95% confidence limits	coefficients in eqs 4 and 5		
(R1,R2) = (2,2)	LiC, BF	28	5.417	K3 = 3.5347	K4 = −0.4749	K5 = 0.03199
				K6 = −0.4749	K7 = 0.03199	K8 = 0.02978
(R1,R2) = (2,3)	LiSi, AlF	25	4.241	K3 = 3.4683	K4 = −0.3786	K5 = 0.02732
				K6 = −0.3423	K7 = 0.02384	K8 = 0.01462
(R1,R2) = (3,3)	NaSi, AlCl	20	6.237	K3 = 3.6358	K4 = −0.2863	K5 = 0.01861
				K6 = −0.2862	K7 = 0.01860	K8 = 0.01062
(C1,C2) = (1,1)	NaLi, RbCs	8	8.301	K0 = 1.2767	K1 = 0.9027	K2 = 0.9027
(C1,C2) = (1,17)	NaF, RbBr	20	5.274	K0 = 0.3855	K1 = 0.8099	K2 = 0.9252
(C1,C2) = (13,17)	BI, TlCl	19	6.975	K0 = 0.3108	K1 = 0.6596	K2 = 0.8627
(C1,C2) = (14,15)	CN, SnP	18	6.752	K0 = 0.2362	K1 = 0.6701	K2 = 0.7413
(C1,C2) = (15,15)	NAs, PBi	12	9.530	K0 = 0.2080	K1 = 0.7063	K2 = 0.7063
(C1,C2) = (15,17)	NF, PBr	7	4.282	K0 = 0.2407	K1 = 0.7163	K2 = 0.7789
(C1,C2) = (16,16)	SO, TeS	14	5.486	K0 = 0.1244	K1 = 0.7589	K2 = 0.7589
(C1,C2) = (17,17)	FCl, BrI	16	5.818	K0 = 0.3555	K1 = 0.7077	K2 = 0.7077

The last two columns of the table present the average of all of the results for each molecule, and their errors, where both are computed using the inverse square of the individual errors as weights. The average error, based on the average of the absolute differences between the predicted values and tabulated values for the molecules having Huber and Herzberg data, is 0.074 Å; if each error is expressed as a percent of the forecast to which it pertains, the average of these errors is 2.94%.

A sample of 25 molecules was collected from those tabulated in Huber and Herzberg but not used for the training of the neural network model or for the least squares smoothing. The absolute differences between (a) the forecasts in Table A, Supporting Information and (b) the tabulated data were divided by the errors of the forecasts (found in the last column of the table). These ratios were sorted into bins with integer upper bounds. Ignoring two bins each containing one obvious outlier, the 23 remaining ratios fall into the first six bins and strongly suggest a Poisson distribution. The distribution has its maximum toward the low side of the bin containing differences between one and two times the errors of the forecasts, and 14 of the 23 differences (61%) have magnitudes less than twice the errors of the forecasts. Details of this and the following analysis may be obtained from the corresponding author.

A sample set of 69 data for 37 molecules was gleaned from articles listed in recent issues of Davis and Eakin.[2] The absolute differences between the forecasts and the tabulated data, divided by the errors of the forecasts, were sorted as before. The most plausible description of the differences is that there are 22 outliers in the bins with upper limits equal to eight or more times the errors of the forecasts; that the 47 remaining absolute differences fall into bins that suggest a Poisson distribution with its maximum in the bin containing differences between one and two times the errors; and that 23 of the differences (49%) have magnitudes less than twice the errors of the forecasts.

The large number of outliers and the lower percentage of differences with magnitudes less than twice the errors of the forecasts illustrate why literature data were not used for the data mining. As elegant and precise as they may seem, and as well as computation and experiment are stated to be in agreement, the internuclear separations reported in any given article require critical comparisons with other contributions before there can be complete confidence in their accuracies.

5. DISCUSSION

There is a second reason data published in the literature since 1979 were not included in the training processes. A heuristic calculation suggests that to have personnel sufficiently competent to distill the journal articles cited in a 25-year collection of the "Berkeley Newsletter" would cost in the neighborhood of $100 000.

Other atlases of globally predicted data for vibration frequencies and ionization potentials of diatomic molecules are in preparation. These efforts have already yielded an interesting phenomenon of possible theoretical interest: certain series of triatomic molecules have, for at least one property and for certain regions of chemical data space, chemical similarity at least as pronounced as do isoelectronic series in many regions of the space.[14,15]

ACKNOWLEDGMENT

The colleagues (Dr. H. Kuhlman and undergraduate students at SAU) who participated in various aspects of this work are too numerous to mention here. Their names are given on pages 185, 186, and 293 of ref 3 and in the bylines and acknowledgments of refs 4, 5, 14, and 15. The reviewers and publisher contributed invaluable comments.

Supporting Information Available: Predictions of neutral diatomic molecule ground-state internuclear separations (Table A). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Huber, K. P.; Herzberg, G. *Constants of Diatomic Molecules*; Van Nostrand Reinhold: New York, 1979. Data for individual molecules are available on-line, courtesy of the National Institute of Standards and Technology, by going to http://www.webbook.nist.gov/chemistry/form-ser.html and selecting the "chemical formula," "Constants of Diatomic Molecules," and "submit."

(2) The Berkeley Newsletter of Molecular Spectra (bimonthly); Davis, S. P., Eakin, D. M., Eds.; Department of Physics, University of California: Berkeley, CA 94720-7300.

(3) Hefferlin, R. *Periodic Systems of Molecules and their Relation to the Systematic Analysis of Molecular Data*; Edwin Mellen Press: Lewiston, NY, 1989; pp 257−277.

(4) Carlson, C.; Gilkeson, J.; Linderman, K.; LeBlanc, S.; Hefferlin, R. Global Forecasting of Data using Least-squares Methods and Molecular Databases: a Feasibility Study using Triatomic Molecules. *Croat. Chem. Acta* **1997**, *770*, 479−508.

(5) Hefferlin, R.; Davis, W. B.; Laing, B. The Learning and Prediction of Triatomic Molecular Data with Neural Networks. In Demidov, V., IAS'97. Murmansk. Proceedings of the Second International Arctic Seminar, Physics and Mathematics, Murmansk State Pedagogical Institute, Murmansk, Russia, 1997, pp 31−36.

(6) Cundari, T. R.; Moody, E. W. Comparison of Neural Networks versus Quantum Mechanics for Inorganic Systems. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 871−875.
(7) Hefferlin, R. Matrix-Product Periodic Systems of Molecules. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 314−317.
(8) Hefferlin, R. *Periodic Systems of Molecules and their Relation to the Systematic Analysis of Molecular Data*; Edwin Mellen Press: Lewiston, NY, 1989; pp 353−380, 387−392.
(9) Gazquez, J. L.; Parr, R. G. *Chem. Phys. Lett*. **1979**, *66*, 419−425.
(10) Nalewajski, R. F.; Thakkar, A. J. Correlations between Average Atomic Numbers and Spectroscopic Constants of Diatomic Molecules. *J. Phys. Chem.* **1983**, *87*, 5361−5367.
(11) Hefferlin, R. The Periodic Systems of Molecules: Presuppositions, Problems, and Prospects. In *Boston Studies in the Philosophy of Science*; Baird, D., Scerri, E., McIntyre, L., Eds.; Kluwer Academic Publishers: in press.
(12) Hefferlin, R. *Periodic Systems of Molecules and their Relation to the Systematic Analysis of Molecular Data*; Edwin Mellen Press: Lewiston, NY, 1989; pp xxiii−xxiv, xxvi−xxxi, 60−164.
(13) A recent URL is http://store.traders.com/v1355prodrev.html.
(14) Cavanaugh, R.; Marsa, R.; Robertson, J.; Hefferlin, R. Adjacent DIM Isoelectronic Molecules and Chemical Similarity among Triatomics. *J. Mol. Struct*. **1996**, *382*, 137−145.
(15) Hefferlin, R.; Thomas Matus, M. Molecular Similarity for Small Species: Refining the Isoelectronic Index. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 484−494.

CI020291Q