Database Mining Using Soft Computing Techniques. An Integrated Neural Network−Fuzzy Logic−Genetic Algorithm Approach

Thomas R. Cundari*[*,†] and Marco Russo[‡]

Department of Chemistry, Computational Research on Materials Institute (CROMIUM), The University of Memphis, Memphis, Tennessee 38152-6060, and Department of Physics−Corpo A−Stanza 18, University of Messina, C. da Papardo, Salita Sperone 31, Villaggio Sant'Agata 98166 (ME), National Institute of Nuclear Physics (INFN) Section of Catania, Corso Italia 57, 95127 (CT), Italy

Received January 30, 2000

Two different soft computing (SC) techniques (a competitive learning neural network and an integrated neural network−fuzzy logic−genetic algorithm approach) are employed in the analysis of a database subset obtained from the Cambridge Structural Database. The chemical problem chosen for study is relevant to the relationship between various metric parameters in transition metal imido (L_nMdNZ, Z = carbon-based substituent) complexes and the chemical consequences of such relationships. The SC techniques confirmed and quantified the suspected relationship between the metal−nitrogen bond length and the metal−nitrogen−substituent bond angle for transition metal imidos: increased metal−nitrogen−carbon angles correlate with shortened metal−nitrogen distances. The mining effort also yielded an unexpected correlation between the NC distance and the MNC angle—shorter NC correlate with larger MNC. A fuzzy inference system is used to construct an MNred−NC−MNC hypersurface. This hypersurface suggests a complicated interdependence among NC, MN_{red}, and the angle subtended by these two bonds. Also, major portions of the hypersurface are very flat, in regions where MNC is approaching linearity. The relationships are also seen to be influenced by whether the imido substituent is an alkyl or aryl group. Computationally, the present results are of particular interest in two respects. First, SC classification was able to isolate an "outlier" cluster. Identification of outliers is important as they may correspond to unreported experimental errors in the database or novel chemical entities, both of which warrant further investigation. Second, the SC database mining not only confirmed and quantified a suspected relationship (MN_{red} versus MNC) within the data but also yielded a trend that was not suspected (NC versus MNC).

INTRODUCTION

As data mining becomes more pervasive for analysis of existing chemical systems and exploitation of this information for the design of novel systems, more research is needed on novel methods for extraction of the maximum information content in databases and database subsets.[1] In data mining one typically seeks to effect a data reduction, e.g., replacement of many data points with representative data clusters. Methods should ideally be automatic and robust, e.g., not requiring extensive statistical pretreatment of the database subset. Another desirable quality for data mining techniques is general applicability to different types of databases and database subsets, for example, bond lengths and bond angles obtained from the Cambridge Structural Database (CSD[2]) as well as drug or catalyst activities from a large combinatorial chemistry study.[3] Methods should be efficient enough to apply to large (either in terms of the number of objects or the dimensionality of the objects) data sets. Finally, one seeks to complement traditional statistical and graphical approaches for data mining by confirming/denying/quantifying suspected trends and identifying those that are not suspected.

In the present research soft computing (SC) techniques[4−6] are evaluated for the analysis of database subsets obtained from the CSD. Specifically, a competitive learning (CL) neural network (NN) is used to classify a 652-object/three-dimension database subset. The database subset is of low dimensionality, but prototypical of datasets obtained from chemical databases, i.e., numerous examples ($\approx 10^2-10^4$ objects) with a few ($\approx 10^0$ characteristics per object) pertinent parameters. The CL classification is augmented by analysis of the same data set using an integrated genetic algorithm−fuzzy logic−neural network tool developed by Russo. The database chosen for the present research is pertinent to the analysis and design of transition metal complexes.

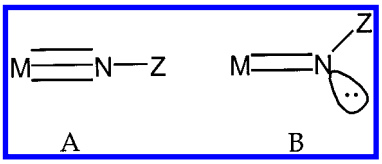
Chemists have long recognized the importance of structural information in providing insight into chemical systems. Traditionally, structural information is obtained from experimental techniques such as X-ray, neutron, or electron diffraction. Recently, software and hardware advances have made it possible to calculate geometric properties (more so for discrete than extended systems) quantitatively, i.e., to a level commensurate with experimental error.[7] Most recently, a "third way" has emerged, which is in some respects a combination of experiment and computation—the use of software to search, classify, and analyze information obtained from a database of experimental structures, of which the Cambridge Structural Database[2] is an excellent example. In its most recent versions the CSD contains close to 200K experimental structures, roughly half of which contain a

* To whom correspondence should be addressed.
† The University of Memphis.
‡ University of Messina.

10.1021/ci0000068 CCC: $20.00
Published on Web 01/26/2001

transition metal. Mining of the CSD has blossomed into a vibrant area of research.[1,8] Typically, the mining of information from database searches is achieved using statistical and graphical approaches. Much less common is the use of soft computing techniques for data mining.[8]

Experimental and computational chemists have extensively probed transition metal (TM) imido ($L_nM{=}NZ$) complexes.[9] Imido complexes have, for example, been implicated as intermediates in nitrogen fixation and have also been shown to effect C−H activation. Of special interest is the nature of correlations between the metal−nitrogen distance and the metal−nitrogen−substituent (Z) angle. Two limiting structures immediately suggest themselves—a linear, triply bonded structure (**A**) and a bent, double bonded structure (**B**). Such analyses were given impetus by the X-ray structure of $Mo(NPh)_2(S_2CNEt_2)_2$ by Haymore and co-workers.[9b] The $Mo(NPh)_2(S_2CNEt_2)_2$ structure shows a large disparity in MoN bond lengths ($\delta r = 0.04$ Å) and $MoNC_{ipso}$ bond angles ($\delta\Theta = 30°$). However, a more recent analogue (Z = Ar = 2,6-diisopropylphenyl) with bulkier substituents possesses two nearly linear ($MNC_{ipso} = 169°$) imido ligands; the equivalent bond lengths (MoN = 1.77 Å) are intermediate to those of $Mo(S_2CNEt_2)_2(NPh)_2$.[9c]



Other variations on the two major themes (**A** and **B**) have been forwarded. High-level ab initio calculations by Cundari[9i] on a wide assortment of TM imidos suggest a linear combination of eight resonance structures is needed to describe the ground-state wave function of transition metal imidos. However, these and other studies do not detract from the essential picture that has suggested a correlation between the MN distance and the MNZ angle.[9] Such musings, apart from the fundamental insight they yield into imido complexes, are pertinent to the rational design of transition metal materials and catalysts. For example, one would expect ligand/substituent combinations that favor **B** to be more desirable in catalyst applications since **B** has a formal metal−nitrogen double bond and a stereochemically active lone pair.

Despite the interest in the **A**/**B** dichotomy, no systematic structural evaluation of the issue has, to our knowledge, been reported. Rankin et al. have reported a study of approximately one dozen polyimido structures.[9d] Given the growth of crystallographic databases, it is worth addressing structural aspects of imido chemistry for a very large assortment of complexes and probing the implications for their bonding, reactivity, and the rational design of imido complexes with tailored chemistry.

METHODS

A search of the Cambridge Structural Database[2] is performed for imido complexes of all types. Complexes in which the imido moiety forms part of a chelating ring are manually deleted. Refinements are employed to ensure the search focused on the most reliable crystal structures: INSIST-ON-COORDS (to search only systems in which fractional atomic coordinates are deposited, ≈88% of the database), INSIST-NO-DISORDER (to search systems with no crystallographic disorder, ≈85% of the database), INSIST-PERFECT-MATCH (to search entries with completely matched chemical and crystallographic connectivities, ≈75% of master database), INSIST-ERROR-FREE (to include only entries whose published bond lengths agree with the recalculated values to within 0.05 Å, ≈98% of the database), and INSIST-NO-POLYMERS (to exclude entries with polymeric bonds in the crystal connectivity, ≈98% of the master database). Statistical outliers (i.e. those structures with values lying more than ±4 standard deviations from mean values of MN_{red} (vide infra), the NX bond length, or the MNC angle) are removed from analysis in order to eliminate structures with unresolved disorder and other errors.

Two imido motifs were retrieved (subject to the constraints above) in the CSD database (**1**, $M{-}N{-}C^3$; **2**, $M{-}N{-}C^4$). Superscripts indicate coordination numbers of the substituent carbons. For the $M{-}N{-}C^3$ dataset, azavinylidenes ($L_nM{=}NdCR_2$), seven of which are found in the CSD, are excluded as these can be easily distinguished from the dominant aryl and vinyl substituents of interest. To facilitate comparison among complexes with different transition metals, the covalent radius of the metal is subtracted from the crystallographically determined M−N bond length to yield the reduced bond length, MN_{red}.

Mining of the imido database is accomplished with soft computing (also referred to as artificial intelligence) techniques, specifically a competitive learning neural network, and an integrated neural network−fuzzy logic−genetic algorithm tool. The latter is implemented into a C++ program by Russo.[10] The CL networks are constructed using the MATLAB software.[11]

Competitive Learning Neural Networks. Competitive learning neural networks are a family of self-organizing networks.[12] The distinctive feature of CL networks is that an input vector **p** (bond lengths and angles in this research) causes a response in the neuron that is closest to it (the "winner") as calculated by the distance between **p** and the input weight matrix **w** of the neuron. A "winner-take-all" strategy is employed, i.e., only the winner neuron outputs 1, while all other neurons output 0 for a particular input vector. Another important feature of CL networks is that neurons that are physically close to each other in the user-defined neuron layer respond to similar input vectors. The Kohonen learning rule[12] is employed; the goal of this learning rule is to have a neuron output a 1 when a vector **p** belongs to its particular cluster and have all other neurons output 0 for that vector **p**. In the present research, a one-dimensional, 20-neuron layer is employed. Training time for a 20-neuron, 100K-epoch simulation is 12.5 min for our 652-by-3 dataset on a 450 MHz PentiumIII personal computer running Windows98.

GEFREX. GEFREX is a genetic fuzzy rule extractor for fuzzy supervised learning.[10] A genetic algorithm is used to optimize the premises of a fuzzy rule; the antecedents in the premise are connected by ANDs (fuzzy minimum). Each antecedent (x_i is A_{ir}) in the premise is described by a Gaussian membership function,

$$\mu_{A_{ir}}(x_i) = \exp(-\gamma_{ir}^2 * (x_i - c_{ir})^2)$$

where x_i is the ith crisp input (bond length or bond angle)

and A_{ir} is a fuzzy set for the rth rule. The genetic coding involves optimization of c_{ir} (center) and γ_{ir} (width) of the Gaussian membership functions (real functions) and an enabling bit (binary), which determines whether the rule is to be employed or not. Training times for GEFREX are less than 1 min for our 652-by-3 dataset on a 350 MHz personal computer.

RESULTS AND DISCUSSION

Carrying out the search outlined above results in 371 M−N−C^3 and 281 M−N−C^4 motifs from 269 and 166 different crystal structures, respectively. For M−N−C^4 systems the dominant (≈90%) substituent was Z = ^{t}Bu, with a greater assortment for M−N−C^3 species (primarily (≈95%) Z = Ph, *p*-tolyl, Ar, 2,6-$C_6H_3Me_2$, mesityl). Therefore, datasets **1** and **2** ostensibly model aryl− and alkyl−imido complexes, respectively. The different imido motifs are combined into a single database subset of 652 objects, each with three dimensions (MN_{red}, NC, and MNC).

Using traditional graphical and statistical approaches on the entire 652-by-3 database subset, it was not possible to discern any trends with respect to the proposed correlations between MN_{red} and MNC. This was the case even if the entire dataset was partitioned using a priori chemical knowledge, e.g., dividing the database into alkyl- and aryl-substituted imidos, partitioning by specific transition metals, and so forth. Hence, we sought to investigate soft computing methods to "cluster" the data and extract chemically meaningful analyses with minimal data preparation.

1. Competitive Learning Neural Networks. Two network types are most often employed for classification problems—competitive learning networks and self-organizing maps (SOM).[5,12] Others have discussed the advantages and disadvantages of each network type, as well as specific architectural differences.[5,12] Our motivation for choosing CL networks over SOMs was due to the fact that work in our laboratory suggests that the latter tend to be more sensitive to the network specifics such as the number and geometry of the nodes and the distance algorithm employed.[13] Competitive learning neural networks are employed in this research for classification of the three metric characteristics—reduced metal−nitrogen bond length (MN_{red}), nitrogen−carbon bond length (NC), and metal−nitrogen−carbon bond angle (MNC)—for each of the 652 objects (imido motifs **1** and **2** combined)—obtained from the CSD search. The data are not normalized; other research in our laboratories suggests that normalization tends to reduce the information content of metric data sets.[13]

Experimentation with different CL simulation details (number of epochs, learning rate, size of the hidden layer, etc.) indicates that classification results are not inordinately sensitive to these parameters. This is encouraging with respect to general utilization of the techniques. The work shown below involves CL networks with a maximum of 20 nodes and 100K epochs. This architecture was employed for 1000 separate CL simulations, each initialized with different, random node-to-node connection weights. The simulations used metric data from the CSD without normalization. From the 1000 simulations the "best" CL network is extracted. The working definition for best is that CL simulation for which the product of the Spearman rank correlation coefficients (ρ_s)[14] for $\langle MN_{red}\rangle$ versus $\langle MNC\rangle$ and $\langle NC\rangle$ versus $\langle MNC\rangle$ (see Figure 1) is the highest. The Spearman formula utilizes not individual values but assigns each value a rank from 1 (lowest) to N (highest), where N is the number of sample data points.[14] The average of MNC, MN_{red}, and NC for each particular cluster defines that cluster's "centroid". In this paper cluster centroids are denoted by angular brackets, viz., $\langle MNC\rangle$, $\langle MN_{red}\rangle$, and $\langle NC\rangle$. It is worth stating that the robustness of the CL approach is suggested by the small variation in ρ_s from one simulation to the next.

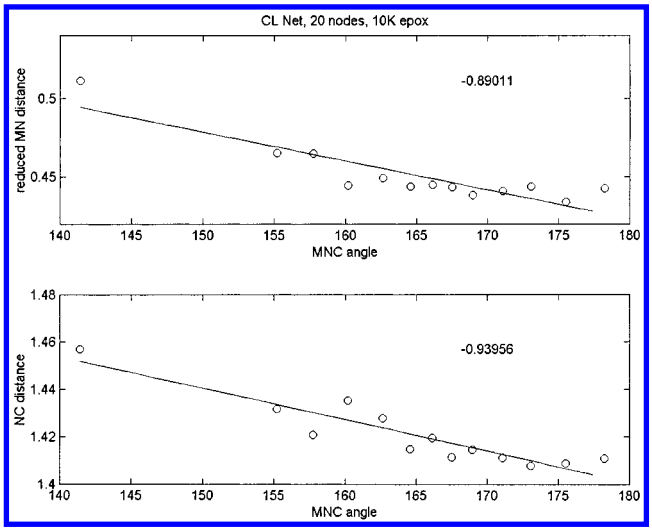


Figure 1. Plot of competitive learning-derived cluster centroids for reduced metal−nitrogen bond length (MN_{red}, Å) versus metal−nitrogen−carbon bond angle (MNC, °), top, and nitrogen−carbon bond length (NC, Å) versus MNC, bottom. See Table 1 for data. The Spearman rank correlation coefficient (ρ_s) is shown in the upper right-hand corner of each graph.

Table 1. Competitive Learning Neural Network Classification of Imido Complexes[a]

⟨MNC⟩ (deg)	⟨MN_{red}⟩ (Å)	⟨NC⟩ (Å)	cluster rank[b]	no.
175.5	0.43	1.41	1	103
173.1	0.44	1.41	2	75
171.1	0.44	1.41	3	75
169.0	0.44	1.41	4	61
178.2	0.44	1.41	5	56
166.1	0.45	1.42	6	50
164.6	0.44	1.41	7	47
160.2	0.44	1.44	8	42
141.4	0.51	1.46	9	35
167.5	0.44	1.41	10	32
162.7	0.45	1.43	11	27
157.7	0.46	1.42	12	26
155.2	0.47	1.43	13	23

[a] Quantities in angular brackets denote the cluster centroids, i.e., the average of the particular characteristic for all objects that belong to a particular cluster. [b] Object clusters are ranked by their membership size (given in the final column) from largest to smallest.

The results for the CL object classification are organized in Table 1 and Figure 1. Table 1 shows the calculated characteristic averages for the objects that comprise each of the thirteen clusters obtained. In the present case, characteristics are metric parameters. Additionally, each cluster is ranked by membership size, i.e., the number of objects that belong to that cluster.

Inspection of Table 1 shows one obvious "outlier" cluster (**9**) with 35 members. Interestingly, cluster **9** is delineated from the others by small MNC angles ($\langle MNC\rangle \approx 141°$), long

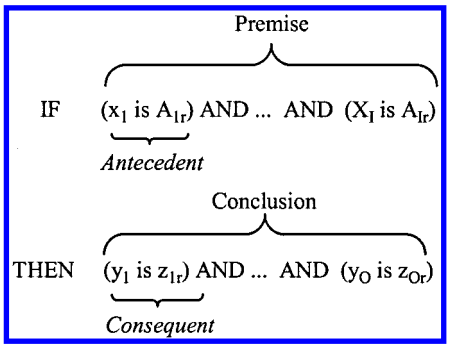
metal−nitrogen bond lengths ($\langle MN_{red} \rangle \approx 0.51$ Å), and long NC distances ($\langle NC \rangle \approx 1.46$ Å).

Figure 1 is a plot of the CL centroids—$\langle MN_{red} \rangle$ versus $\langle MNC \rangle$ and $\langle NC \rangle$ versus $\langle MNC \rangle$ for the thirteen clusters listed in Table 1. As with Table 1, outlier cluster **9** is clearly visible in Figure 1. This is the reason the Spearman scheme is used, since it is less sensitive to outliers than the Pearson correlation coefficient. Figure 1 shows a substantial ($\rho_s \approx -0.89$) correlation between $\langle MN_{red} \rangle$ and $\langle MNC \rangle$. The negative sign of ρ_s indicates that as the MNC angle approaches linear, the MN_{red} distance decreases, confirming and quantifying the long-suspected relationship of the metrics.[9] Equally as interesting is that the CL analysis yields an unexpected relationship between $\langle MNC \rangle$ and $\langle NC \rangle$ ($\rho_s \approx -0.94$).

2. Integrated Neural Network−Fuzzy Logic−Genetic Algorithm. Each SC technique (genetic algorithms (GAs), neural networks (NNs), and fuzzy logic (FL)) has advantages and disadvantages with respect to time complexity and amount of a priori information required. In general, the amount of a priori information needed increases GA < NN < FL, while time complexity increases in the opposite order (FL < NN < GAs).[10] Russo has shown that an SC modeling scheme that merges these techniques can accentuate their advantages while reducing the disadvantages. This methodology was originally implemented in the FuGeNeSys tool,[10a] and more recently as part of the GEFREX program.[10b] The main characteristic of GEFREX is its genetic nature. A neural-based operator is introduced to improve performance with regard to learning speed and error. GEFREX is able to extract fuzzy knowledge in a fully automatic supervised manner and works with approximation, classification, and time series prediction tasks.

In this research, GEFREX is employed to model the TM imido database subset. The objective is to derive a compact fuzzy inference system (FIS) and to use this FIS to simultaneously relate MN_{red} and NC with MNC; i.e., we want to extract from the database the function MNC(MN_{red},NC).

2.1. Fuzzy Knowledge Representation and Inferential Method. The generic *r*th fuzzy rule extracted by GEFREX has the following form:



Each premise contains a maximum of *I* antecedents and each conclusion at most *O* consequents. *I* and *O* are the number of inputs and outputs, respectively. For the present research $I = 2$ (MN_{red}, NC) and $O = 1$ (MNC).

In the generic antecedent (x_i is A_{ir}) there is the *i*th crisp input x_i and the fuzzy set (FS) A_{ir} that is generally different from all the others; e.g., the membership function (MF) of A_{ir} can have a Gaussian, triangular, or trapezoidal shape. All the connectors in the premise are ANDs. The algebraic minimum or the product operator is used for this operator. In the conclusion there are no FSs but only singletons (z_{or}).

Table 2. GEFREX Derived Fuzzy Inference System for Imido Complexes[a]

rule (Å^{-1})	$\gamma(MN_{red})$ (Å)	$c(MN_{red})$ (Å^{-1})	γ(NC) (Å)	c(NC) (deg)	z(MNC)
1	44.12	1.411	88.80	0.5090	170.9
2	30.19	1.331	91.00	0.4885	167.2
3	27.81	1.432	57.64	0.4098	171.5
4	79.17	1.468	21.00	0.3613	172.5
5	29.49	1.517	19.00	0.5829	131.1
6	27.81	1.511	36.51	0.4329	164.0

[a] The GEFREX-optimized input (i.e., MN_{red} and NC) widths (γ) and centroids (c) for the Gaussian membership functions (μ) used for determination of the six-rule fuzzy inference system. Outputs are treated as fuzzy singletons. Rules are of the form given in eq 1. The GEFREX simulations employed half of the 672 object imido data set for training and half for testing.

Two different versions of the defuzzification method have been implemented.[10] The first defuzzification method is

$$y_o = \frac{\sum_{r=1}^{R} \theta_r z_{or}}{\sum_{r=1}^{R} \theta_r} \tag{1}$$

where R is the total number of rules and θ_r is the degree of truth of the *r*th rule. This method is a weighted mean (WM) defuzzification and is similar to the Yager method.[15] If all θ_r are zero, then all the outputs are zero. The second defuzzification method is simpler. It is the so-called weighted sum (WS):

$$y_o = \sum_{r=1}^{R} \theta_r z_{or} \tag{2}$$

The results of the GEFREX simulations are shown in Table 2 and Figure 2. Gaussian membership functions are chosen. The TM imido database is divided in two equal parts. The first half is used as learning and the second half as testing. Several learning phases are then executed by changing the number of rules and the defuzzification method. Each simulation is evaluated by considering only the testing error. The best results found correspond to WM defuzzification and six rules. Average angle differences of 5.8 and 6.5° are found for the learning and testing sets, respectively.

Table 2 tabulates the optimized width (γ) and centroid (c) that define the Gaussian membership functions for both inputs (MN_{red} and NC) and the MNC singletons. Hence, the fuzzy rules are of the form depicted in Figure 2, where x_1 and x_2 are specific values of the reduced metal−nitrogen and nitrogen−carbon distances, respectively. For example, rule 5 can be written in the following way:

IF (MN_{red} is high) AND (NC is high) THEN (MNC = 131.1°) (3)

Implicit in the CL analysis in Figure 1 is an "uncoupling" of the two bond lengths with respect to their influence on

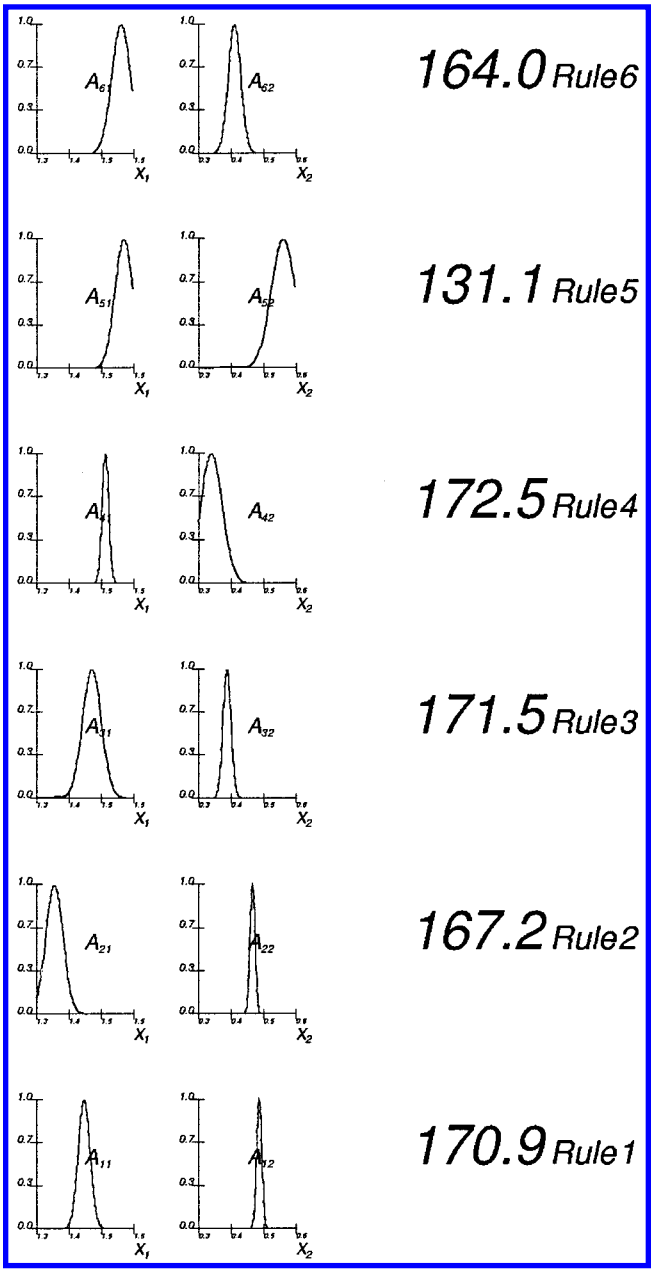


Figure 2. Graphical representation of GEFREX-derived rules that comprise the fuzzy inference system obtained from the imido database subset. These rules (see also Table 2) are used to generate Figure 3.

MNC. It is thus desirable to address their combined effect on the bond angle. The GEFREX-derived FIS is used to generate an MN_{red}–NC−MNC hypersurface. A contour plot showing the outcome of this effort is given in Figure 3 for the chemically relevant portion of parameter space: MN_{red} = 0.40−0.60 Å and NC = 1.35−1.50 Å. To highlight chemical aspects of the SC data mining, (MN_{red},NC) pairs for all 652 data points are overlaid on top of the FIS-derived hypersurface and differentiated into motifs **1** (aryl imidos, "◆" in Figure 3) and **2** (alkyl imidos, "+" in Figure 3). Several features are seen in Figure 3 and are germane to TM imido research.

The MNC surface is very flat for MN_{red} < 0.53 Å and NC < 1.41 Å, values that correspond to the 97th percentile and median of MN_{red} and NC, respectively. Roughly 50% of the imido data fall within this region.[16] As is evident from Figure 3, this large subset of the data is populated primarily by aryl imidos immediately suggesting a greater degree of structural flexibility for alkyl than aryl imidos.

In general, MNC is less sensitive to NC than MN_{red}, particularly for the small (<0.45 Å) MN_{red} values that comprise the majority of reported imido crystal structures. However, as the metal−imido bond lengthens (MN_{red} > 0.48 Å), greater sensitivity of MNC toward NC is predicted, particularly for NC > 1.45 Å. Figure 3 suggests that small MNC angles are observed when (a) MN_{red} is very large (>0.55 Å) regardless of the NC distance or (b) MN_{red} > 0.48 Å and NC > 1.45 Å. Inspection of Figure 3 indicates that imidos with NC > 1.45 Å are almost exclusively alkyl imidos. Indeed, alkyl imidos **2** have significantly longer NC bond lengths (1.46 ± 0.02 Å) on average than aryl imidos **1** (NC = 1.39 ± 0.02 Å). A difference of 0.07 Å is much larger than would be expected by a change in hybridization (without a change in NC bond order) from N−C_{sp^3} (**2**) to N−C_{sp^2} (**1**), which would be 0.01−0.02 Å.[17]

Hence, the metal−nitrogen and nitrogen−carbon bonds are both seen to influence (or are influenced by) the bond angle they subtend. A degree of caution is warranted as Figure 3 makes it obvious that the number of examples with MN_{red} > 0.48 Å is relatively sparse. Hence, the synthesis and structural characterization of complexes with relatively long metal−imido bonds (perhaps through the use of π-loading[9a]) and long NC bond lengths is warranted. With respect to the latter design criterion, the SC analysis suggests that alkyl imidos are a more profitable target.

SUMMARY AND CONCLUSION

This paper reports the analysis of a database subset obtained from the Cambridge Structural Database. Two different soft computing techniques are employed: a competitive learning neural network and an integrated neural network−fuzzy logic−genetic algorithm approach. For the former the mining effort is undertaken as a classification problem, while the latter approaches the data mining as an exercise in function approximation. The specific chemical problem chosen for study is relevant to the relationship between various metric parameters for transition metal imido complexes and the chemical consequences of such relationships. Several important conclusions have been reached, the most important of which are summarized here.

(1) The soft computing techniques confirmed and quantified the suspected trends between the metal−nitrogen bond length and the metal−nitrogen−substituent bond angle for TM imidos. Increased metal−nitrogen−carbon angles are seen to correlate with shortened metal−nitrogen distances, Figure 1 (top). Hence metals, ligands, and substituents that favor a double bonded structure (**B**) would be expected to be favorable from the point of view of catalyst design.

The database analysis suggests this can be accomplished through the synthesis of imido complexes with long metal−nitrogen and long nitrogen−carbon bond lengths. For the latter, alkyl groups would seem to be the desired substituents. In terms of yielding high MN_{red}, this might be possible through π-loading,[9a] or otherwise weakening the metal−nitrogen π bond, for example, through targeting late, low-valent transition metals.

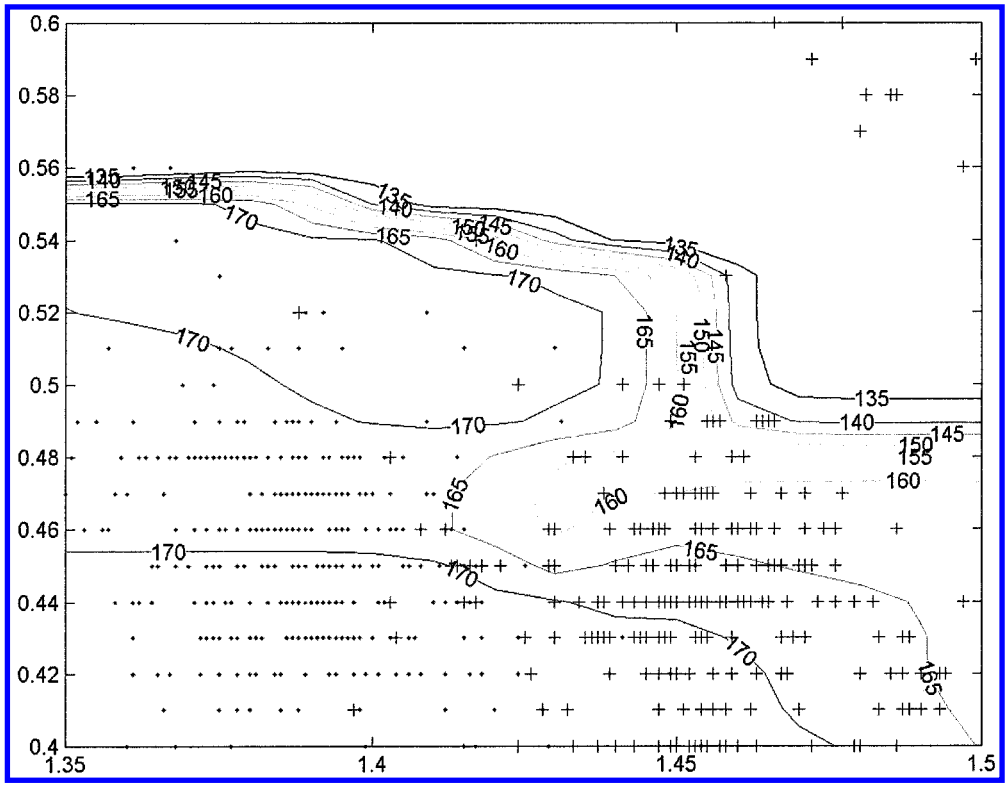


Figure 3. Contour plot of MN_{red} (ordinate) versus NC (abscissa) versus MNC. The grid size used for generating the contour surface is 0.01 Å for both MN_{red} and NC. Contour levels for MNC (deg) are indicated. Plus signs and diamonds denote MN_{red}, NC pairs obtained from the alkyl− and aryl−imido portion of the data, respectively.

(2) The mining effort also yielded an unexpected correlation between the nitrogen−carbon distance and the MNC bond angle—shortened NC correlate with increased MNC, Figure 1 (bottom). The chemical causes of the correlation are not immediately certain. Two hypotheses seem plausible. First, large MNC angles (structure **A**) imply sp hybridization of the imido nitrogen as opposed to sp^2 hybridization for bent imidos (**B**). Hence, the greater s character of the former will engender shorter N−Z bond lengths. Second, for the subset of aryl−imido complexes (more than half of the 672 data objects) a larger angle may induce shorter bonds by accentuating resonance between the metal−nitrogen multiple bond and the arene ring. Quantum calculations on an assortment of aryl−imido complexes show extensive delocalization across the MNAryl fragment.[18] Shorter NC bond distances (≈0.07 Å on average) for aryl imidos versus alkyl imidos are consistent with significant delocalization between the MN_{imido} and aryl moieties. This suggests that a profitable route to rational modification of aryl−imido complexes may be achieved by the utilization of heterosubstituted aryl−imido complexes since the vast majority of structurally characterized aryl−imido complexes possess phenyl or alkyl−phenyl (Ar, Ar′, mesityl, and *p*-tolyl) substituents.

(3) Using a neural network−fuzzy logic−genetic algorithm scheme for function approximation shows the interrelationships among the three metrics in more detail. A fuzzy inference system is used to derive an MN_{red}−NC−MNC hypersurface, Figure 3. For example, major portions of the hypersurface are very flat, in regions where MNC is approaching linearity (>170°). It is particularly noteworthy that the flattest portion of the MNC hypersurface is primarily populated by aryl−imido complexes, suggesting that such species demonstrate a lesser degree of structural flexibility than alkyl imidos.

Hypersurfaces constructed from pertinent experimental metrics such as those in Figure 1 (two-dimensional) and Figure 3 (three-dimensional) have been widely used in structural database analysis to construct experimental trajectories and derive dynamic information from static crystal structures.[19] These "Bürgi-Dunitz" hypersurfaces are almost exclusively constructed from single data points. The use of SC techniques for data reduction of numerous objects to representative clusters and the determination of average cluster metrics represent an exciting approach for the construction of Bürgi-Dunitz hypersurfaces from very large database subsets. The use of data derived from numerous data points should reduce issues that may arise from outlier or incorrect experimental structures.

From a computational point of view, the present results are exciting in two respects. First, the neural network classification was able to isolate an "outlier" cluster. Identification of outliers is important since they may correspond to unreported experimental errors in the database, or chemical entities with novel properties. In either case, such entities would be worthy of further investigation using quantum mechanical methods. Second, the soft computing-based database mining effort not only confirmed and quantified a suspected relationship (MN versus MNC) within the data but also yielded a trend that was not suspected (NC versus MNC). From this perspective, soft computing techniques seem an exciting option for mining of chemical databases and database subsets. Ideally, one can envision the implementation of SC techniques earlier in the mining process (e.g., to aid in extraction of a subset from a database

as opposed to analysis alone). Research directed toward this end is now under way in our laboratories.

ACKNOWLEDGMENT

T.R.C. gratefully acknowledges the National Science Foundation (NSF) for their support through Grants CHE-9614346 and CHE-9983665. Some calculations employed the Computational Chemistry Resource at The University of Memphis, funded by Grant CHE-9708517 from the NSF Chemical Research and Instrumentation Facilities program. The imido database subset was collected while T.R.C. was a Visiting Fellow at the University of Bristol (U.K.); T.R.C. acknowledges The University of Memphis, the College of Arts & Sciences, and the UM Chemistry Department for a Professional Development Assignment. T.R.C. further acknowledges Prof. A. Guy Orpen (Bristol) and his group for their hospitality and the NSF Office of International Programs for making the visit possible through Grant CHE-9802675.

REFERENCES AND NOTES

(1) For a representative assortment of structural database analyses in transition metal chemistry see the following papers and the references cited therein. (a) Orpen, A. G. Structural Systematics in Molecular Inorganic Chemistry. *Chem. Soc. Rev.* **1993**, 191−197. (b) Orpen, A. G. Structural Systematics. 6. Apparent Flexibility of Metal Complexes in Crystals. Martin, A. *J. Am. Chem. Soc.* **1996**, *118*, 1464−1470, and earlier works in this series. (c) Braga et al. (Braga, D.; Grepioni, F.; Baradha, K.; Desiraju, G. R. Agostic Interactions in Organometallic Compounds. A Cambridge Structural Database Study. *J. Chem. Soc., Dalton Trans.* **1996**, 3925−3930) discuss the use of the CSD to analyze *intramolecular*, agostic interactions for organo-TM complexes. (d) Trnka, T.; Parkin, G.A Survey of Terminal Chalcogenido Complexes of the Transition Metals: Trends in their Distribution and the Variation of their M=E Bond Lengths. *Polyhedron* **1997**, *16*, 1031−1045.

(2) Allen, F. H.; Kennard, O. 3D Search and Research using the Cambridge Structural Database. *Chem. Des. Autom. News* **1993**, *8*, 31−37.

(3) Representative applications of combinatorial chemistry (*Combinatorial Chemistry: Synthesis and Applications*; DeWitt, S. H., Czarnick, A. E., Eds.; American Chemical Society: Washington, D. C., 1997) in catalysis are as follows: Gennari, C.; Ceccarelli, S.; Piarulli, U.; Montalbetti, C. A. G. N.; Jackson, R. F. W. Investigation of a New Family of Chiral Ligands for Enantioselective Catalysis via Parallel Synthesis and High-Throughput Screening. *J. Org. Chem.* **1998**, *63*, 5312−5313. Sigman, M. S.; Jacobsen, E. N. Schiff Base Catalysts for the Asymmetric Strecker Reaction Identified and Optimized from Parallel Synthetic Libraries. *J. Am. Chem. Soc.* **1998**, *120*, 4901−4902.

(4) Judson, R. Genetic Algorithms and Their Use in Chemistry. In *Reviews in Computational Chemistry*; Boyd, D. B., Lipkowitz, K. B., Eds.; VCH: New York, 1997; Vol. 10, pp 1−73.

(5) (a) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH: Weinheim, Germany, 1993. (b) Sumpter, B. G.; Getino, C.; Noid, D. W. Theory and Applications of Neural Computing in Chemical Science. *Annu. Rev. Phys. Chem.* **1994**, *45*, 439−481.

(6) *Fuzzy Logic in Chemistry*; Rouvray, D., Ed.; Academic Press: San Diego, 1997.

(7) For specific discussions concerning the application of computational chemistry to transition metal complexes see the following works. (a) Benson, M. T.; Cundari, T. R.; Lutz, M. L.; Sommerer, S. O. Effective Core Potential Approaches to the Chemistry of the Heavier Elements. *Reviews in Computational Chemistry*; VCH: New York, 1996; Vol. 8, pp 145−202. (b) Frenking, G.; Antes, I.; Bohme, M.; Dapprich, S.; Ehlers, A. W.; Jonas, V.; Neuhaus, A.; Otto, M.; Stegmann, R.; Veldkamp, A.; Vyboishchikov, S. F. Pseudopotential Calculations of Transition Metal Compounds: Scope and Limitations. *Reviews in Computational Chemistry*; VCH: New York, 1996; Vol. 8, pp 63−144. (c) Ziegler, T. Density Functional Theory as a Practical Tool in Studies of Organometallics Energetics and Kinetics. Beating the Heavy Metal Blues with DFT. *Can. J. Chem.* **1995**, *73*, 743−761.

(8) Beyreuther et al. discuss the application of neural networks to the conformational analysis of organometallic TM complexes. Conformation of tripod metal templates in $CH_3C(CH_2)PPh_2)_3ML_n$ (n = 2, 3): Beyreuther, S.; Hunger, J.; Huttner, G.; Mann, S.; Zsolnai, L. Neural Networks in Conformational Analysis. *Chem. Ber.* **1996**, *129*, 745−757.

(9) (a) Wigley, D. E. Organoimido Complexes of the Transition Metals. *Prog.* Inorg. Chem. **1995**, *42*, 239−482. (b) Haymore, B. L.; Maatta, E. A.; Wentworth, R. D. A Bisphenylnitrene Complex of Molybdenum with a Bent Nitrene Ligand. Preparation and Structure of *cis*-Mo(N(C_6H_5))$_2$($S_2CN(C_2H_5)_2$)$_2$. *J. Am. Chem. Soc.* **1979**, *101*, 2063−2068. (c) Coffey, T. A.; Forster, G. D.; Hogarth, G.; Sella, A. A Steric Preference for Linear versus Bent Imido Ligation? Synthesis and X-ray Crystal Structure of [Mo(NAr)$_2$(edtc)$_2$] (Ar = 2,6-^{i}Pr$_2$$C_6H_3$; edtc = S_2CNEt_2) Containing Two Linear Imido Moieties. *Polyhedron* **1993**, *12*, 2741−2743. (d) Rankin, D. W. H.; Robinson, H. E.; Danopoulos, A. A.; Lyne, P. D.; Mingos, D. M. P.; Wilkinson, G. Molecular Structure of Tetrakis(*tert*-butylimido)osmium(VIII), Determined in the Gas Phase by Electron Diffraction. *J. Chem. Soc., Dalton Trans.* **1994**, 1563−1569. (e) Parkin, G.; van Asselt, A.; Leahy, D. J.; Whinnery, L.; Hua, N. G.; Quan, R. W.; Henling, L. M.; Schaefer, W. P.; Santarsiero, B. D.; Bercaw, J. E. Oxo-Hydrido and Imido-Hydrido Derivatives of Permethyltantalocene. Structures of (η^5-C_5Me_5)$_2$Ta(=O)H and (η^5-C_5Me_5)$_2$Ta(=NC_6H_5)H: Doubly or Triply Bonded Tantalum Oxo and Imido Ligands. *Inorg. Chem.* **1992**, *31*, 82−85. (f) Schofield, M. H.; Kee, T. P.; Anhaus, J. T.; Schrock, R. R.; Johnson, K. H.; Davis, W. M. Osmium Imido Complexes: Synthesis, Reactivity, and SCF−Xα−SW Electronic Structure. *Inorg. Chem.* **1991**, *30*, 3595−3604. (g) Green, M. L. H.; Hogarth, G.; Konidaris, P. C.; Mountford, P. Interconversion of Oxo and Imido Ligands at a Dimolybdenum Centre: Molecular and Electronic Structure of [{Mo(η-C_5H_4Me)(NPh)(μ-NPh)}$_2$]. *J. Chem. Soc., Dalton Trans.* **1990**, 3781−3787. (h) Jorgensen, K. A. MO Explanation of the "Unexpected" Structure of (η^5-C_5Me_5)$_2$Ta(=NC_6H_5)H. *Inorg. Chem.* **1993**, *32*, 1521−1522. (i) Cundari, T. R. Transition Metal Imido Complexes. *J. Am. Chem. Soc.* **1992**, *114*, 7879−7888. (j) Gibson, V. C.; Marshall, E. L.; Redshaw, C.; Clegg, W.; Elsegood, M. R. J. Bent versus Linear Imido Ligands in Five-Coordinate Molybdenum Complexes. *J. Chem. Soc., Dalton Trans.* **1996**, 4197−4199. (k) Gibson, V. C.; Clegg, W.; Dwyer, P. M.; Elsegood, M. R. J.; Bell, A.; Marshall, E. L. Novel Bis(imido) Complexes of Molybdenum(VI): Precursors to New Alkene Metathesis Catalysts. *J. Chem. Soc., Chem. Commun.* **1994**, 2247−2248. (l) Bradley, D. C.; Hodge, S. R.; Runnacles, J. D.; Hughes, M.; Mason, J.; Richards, R. L. Nitrogen Nuclear Magnetic Resonance Spectroscopy as a Probe of Bonding, Bending and Fluxionality of the Imido Ligand. *J. Chem. Soc., Dalton Trans.* **1992**, 1663−1668.

(10) (a) Russo, M. FuGeNeSys−A Fuzzy Genetic Neural System for Fuzzy Modeling. *IEEE Trans. Fuzzy Syst.* **1998**, *6*, 373−388. (b) Russo, M. Genetic Fuzzy Learning. *IEEE Trans. Evol. Comput.* **2000**, *4*, 259−273.

(11) *Neural Network Toolbox*, version 3; MATLAB, The MathWorks, Inc.: Natick, MA, 1998.

(12) See ref 5a (pp 83−87) for an expanded discussion of competitive learning neural networks.

(13) Cundari, T. R.; Deng, J.; Pop, H. F.; Sârbu, C. Soft Computing Techniques for Mining Structural Databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1052−1061.

(14) Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*; McGraw-Hill: New York, 1956. Spearman's formula for calculation of the rank correlation coefficient is $\rho_s = 1 - (6(\Sigma d^2)/N(N^2 - 1))$, where N is the sample size and d is the difference in rank of corresponding data pairs.

(15) Figueiredo, M.; Gomides, F.; Rocha, A.; Yager, R. Comparison of Yager's Level Set Method for Fuzzy Logic Control with Mamdani and Larsen Methods. *IEEE Trans. Fuzzy Syst.* **1993**, *1*, 156−159.

(16) There is another flat region above the line ($MN_{red} = 1.12-0.40 \times NC$), although few data points exist in this region of metric parameter space.

(17) Allen, F. H.; Kennard, O.; Watson, D. G.; Brammer, L. G.; Orpen, A. G.; Taylor, R. *J.* Tables of Bond Lengths determined by X-ray and Neutron Diffraction. Part 1. Bond Lengths in Organic Compounds. *Chem. Soc., Perkin Trans. 2* **1987**, S1−S19.

(18) Cundari, T. R. Unpublished results.

(19) (a) Dunitz, J. D.; Bürgi, H. B. *Structure Correlation*; VCH: Weinheim, Germany, 1994. (b) Bürgi, H. B., Dunitz, J. D. From Crystal Statics to Chemical Dynamics. *Acc. Chem. Res.* **1983**, *16*, 153−161.

CI0000068