



## Computational structure–activity relationship analysis of non-peptide inducers of macrophage tumor necrosis factor- $\alpha$ production

Andrei I. Khlebnikov<sup>a,\*</sup>, Igor A. Schepetkin<sup>b</sup>, Liliya N. Kirpotina<sup>b</sup>, Mark T. Quinn<sup>b,\*</sup>

<sup>a</sup> Department of Chemistry, Altai State Technical University, Barnaul 656038, Russia

<sup>b</sup> Department of Veterinary Molecular Biology, Montana State University, Bozeman, MT 59717, USA

### ARTICLE INFO

#### Article history:

Received 16 June 2008

Revised 23 August 2008

Accepted 30 August 2008

Available online 5 September 2008

#### Keywords:

Tumor necrosis factor  $\alpha$

Macrophage

Atom pairs

Molecular descriptors

Structure–activity relationship analysis

### ABSTRACT

Previously, we screened a series of arylcarboxylic acid hydrazide derivatives for their ability to induce macrophage tumor necrosis factor  $\alpha$  (TNF- $\alpha$ ) production and identified 16 such compounds. In the present study, we evaluated 23 additional arylcarboxylic acid hydrazides and found that seven of these compounds also induced macrophage TNF- $\alpha$  production, representing novel compounds with this activity. The total set of active compounds was then used for computational structure–activity relationship (SAR) analysis to further optimize lead molecules. A sequence of (1) linear discriminant analysis, (2) classification tree analysis with linear combination, and (3) univariate splits based on atom pair descriptors led to the derivation of SAR rule-based algorithms with fitting accuracy of 96.5%, 91.9%, and 84.9%, respectively. The SAR rules obtained from classification tree analysis with univariate splits, which was based on three atom pair descriptors only, revealed that the main factors influencing agonist activity of arylcarboxylic acid hydrazide derivatives were the presence of a methyl or trifluoromethyl group in the benzene ring attached to the furan moiety, an alkoxy group in the aromatic ring near the methylenehydrazide linker, and two or more halogen atoms (chlorine or bromine) on one side of the dumb-bell-shaped hydrazide molecule opposed by an aromatic moiety on the opposite side of the molecule. Thus, these rules represent a relatively simple classification approach for de novo design of small-molecule inducers of macrophage TNF- $\alpha$  production.

© 2008 Elsevier Ltd. All rights reserved.

### 1. Introduction

Tumor necrosis factor  $\alpha$  (TNF- $\alpha$ ) is a key cytokine that contributes to immune and inflammatory reactions and is important for both innate and adaptive immunity.<sup>1</sup> Currently, a significant effort is focused on the development of anti-TNF- $\alpha$  agents as therapeutics for treatment of chronic inflammatory conditions, such as rheumatoid arthritis and inflammatory bowel disease.<sup>2</sup> However, TNF- $\alpha$  is also well known for its ability to induce apoptosis of tumor cells, resulting in tumor necrosis, and use of TNF- $\alpha$  in cancer treatment has been pursued.<sup>3</sup> Unfortunately, the clinical use of TNF- $\alpha$  has been limited due to its proinflammatory activity.<sup>4</sup> On the other hand, stimulation of endogenous TNF- $\alpha$  production is still considered a reasonable approach in tumor biotherapy, and several compounds have been found to induce TNF- $\alpha$ , inhibit tumor blood flow, and cause necrosis in experimental tumors.<sup>5</sup> Indeed, a number of small-molecule cytokine inducers have been identified and characterized for their ability to stimulate TNF- $\alpha$

production. For example, both natural and synthetic agents with antimicrobial and antitumor properties, such as imidazoquinolines and taxanes, have been shown to induce a broad range of cytokines in cell culture and/or in vivo.<sup>6,7</sup> Recently, we identified several small-molecule *N*-formyl peptide receptor agonists that potently induced TNF- $\alpha$  production in murine and human macrophages.<sup>8</sup> Interestingly, these compounds all contained an arylcarboxylic acid hydrazide core structure, which is distinct from other known inducers of TNF- $\alpha$  production.

Our analysis of arylcarboxylic acid hydrazides showed that individual ring substituents had significant impact on the potency of these derivatives for inducing macrophage TNF- $\alpha$  production,<sup>8</sup> suggesting that further structure–activity relationship (SAR) analysis of these compounds would contribute to our understanding of their mechanism of action and could lead to the development of additional compounds with enhanced efficacy. Indeed, SAR and quantitative SAR (QSAR) models have been instrumental in understanding molecular mechanisms of action of receptor agonists and antagonists, directing their design, and in virtual screening.<sup>9</sup> To date, non-computational SAR analysis has been performed for a series of taxoids<sup>10,11</sup>; however, there are currently no reported computational SAR models for small-molecule inducers of TNF- $\alpha$  production.

\* Corresponding authors. Tel.: +7 3852 245513/522436; fax: +7 3852 367864 (A.I. Khlebnikov); tel.: +1 406 994 4707; fax: +1 406 994 4303 (M.T. Quinn).

E-mail addresses: [aikh@chem.org.ru](mailto:aikh@chem.org.ru) (A.I. Khlebnikov), [mquinn@montana.edu](mailto:mquinn@montana.edu) (M.T. Quinn).

While a variety of molecular parameters can be used in the computational methods for (Q)SAR analysis,<sup>12,13</sup> some of these parameters are complex physicochemical or geometrical descriptors whose calculation is associated with difficulties due to molecular flexibility and inadequate sampling of conformational space. In contrast, topological indices (i.e., 2D descriptors) obtained from the structural formula of a compound are very attractive because of their simplicity. Recently, we developed an improved approach to SAR methodology based on atom pair descriptors in combination with classical physicochemical and geometrical descriptors and showed that this methodology can detect specific combinations of substructure patterns that confer high or low inhibitory activity against neutrophil elastase.<sup>14</sup> Here, we utilized a similar approach for computational SAR analysis of a large group of arylcarboxylic acid hydrazides, including our previously reported derivatives<sup>8</sup> and several novel analogs identified here in further screening. These studies provide further optimization of these molecules as lead compounds that can induce macrophage TNF- $\alpha$  production and also provide clues to the molecular features required for agonist activity.

## 2. Results and discussion

### 2.1. Identification of novel TNF- $\alpha$ inducers and selection of the molecular set

Previously, we screened a series of arylcarboxylic acid hydrazide derivatives for their ability to induce macrophage tumor necrosis factor  $\alpha$  (TNF- $\alpha$ ) production and found that 16 compounds induced production of modest-to-high levels of TNF- $\alpha$  by murine and human macrophages.<sup>8</sup> Structures of these compounds and their activity, expressed as fold-increase (FI) in macrophage TNF- $\alpha$  production above solvent control, together with the inactive arylcarboxylic acid hydrazides that we evaluated previously are shown in Table 1 (compounds **1**, **8–10**, **23–50**, and **52–82**). FI was used to normalize the activity for experiment-to-experiment variations observed in background due to solvent (DMSO) alone. Variations in background activity are likely due to differences in batches of our cultured macrophages, as it is clear that the number of passages affects cell activity, and newer batches of cells exhibited much higher stimulated activity as well as much higher background activity. Since FI represents relative activity above background, use of FI values allowed us to compare results from a number of experiments regardless of background, and average FI from three independent experiments are provided.

To increase the molecular data set, we selected 23 additional arylcarboxylic acid hydrazide derivatives and evaluated their ability to stimulate TNF- $\alpha$  production. As shown in Table 1, we identified seven additional novel compounds with varying levels of activity (compounds **2–7** and **51**). Derivatives of nicotinic acid (compound **2**) and isonicotinic acid (compound **51**) were the most active, inducing similar levels of TNF- $\alpha$  that were induced by control LPS (50 ng/ml) and the most potent of our previously identified compounds (Fig. 1). Activation of macrophage TNF- $\alpha$  production was not due to endotoxin contamination, since analysis of compounds **2** and **51** for endotoxin using a limulus amoebocyte lysate assay showed that these compounds contained no endotoxin (below detection limit; data not shown). Furthermore, treatment with the additional compounds, which included 7 active compounds (**2–7** and **51**) and 16 inactive compounds (**11–22** and **83–86**) from our set, had no effect on cell viability in J774.A1 macrophages, indicating lack of cytotoxicity at concentrations  $\leq 50$   $\mu$ M (data not shown).

In SAR studies a compound set under investigation is conventionally split into two or more classes (active, moderately active,

low-active, non-active, etc.) rather than using individual activities. This allows formulation of more or less simple SAR classification rules, in contrast to a QSAR study where initial numerical values of activity are used (e.g., FI). For SAR analysis here, the total set of the arylcarboxylic acid hydrazide derivatives (compounds **1–86**) was divided into two activity classes based on their experimentally determined activity. Compounds that induced macrophage TNF- $\alpha$  production (FI  $\geq 2$ ) were classified as 'Active' (23 compounds), whereas inactive derivatives were placed in the non-active group labeled 'NA' (63 compounds).

### 2.2. Descriptors

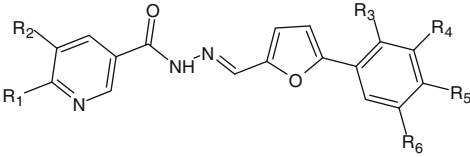
Atom pairs were automatically generated from bond connectivity of the arylcarboxylic acid hydrazides and are specified in terms of types of the two atoms in a pair separated by the number of chemical bonds in the structural formula.<sup>15</sup> As described previously,<sup>14</sup> we used the atom type names from MM+ force field, as implemented in HyperChem. According to this scheme, specific atom pairs are defined as T1\_D\_T2, where T1 and T2 are the atom types assigned by HyperChem, and D is the number of chemical bonds in the shortest path between the two atoms (see Section 4). HyperChem output in a HIN file format was entered directly into our CHAIN program, which generated all possible atom pairs and frequencies of their occurrence in each of the 86 hydrazides. These frequencies were considered as values of the corresponding atom pair descriptors, and examples of atom pairs are shown in Figure 2. Note that atom pair descriptors are easily interpretable in terms of standard chemical formulae. For example, BR\_11\_CA indicates the simultaneous presence of a bromine atom and an aromatic ring in the opposite sides of a molecule (see Fig. 2). It should be noted that, although atom naming was taken from MM+ force field, performing MM+ molecular mechanics optimization itself is not necessary because only bond connectivity, but not geometry, is important for the atom pair calculation.

In total, 836 unique atom pairs were generated for all 86 hydrazides, and a histogram of the number of atom pairs with different bond distances is presented in Figure 3A. Note that the histogram has two maxima at 5 and 10–11 chemical bonds, which is in agreement with the dumbbell shape of most compounds in our set. Indeed, all of the molecules contain two bulky moieties connected by the hydrazide linker. Hence, the relatively 'short' atom pairs originated from the same moiety, as well as the much 'longer' atom pairs representing atoms in the two different moieties prevail in the total number of 836 descriptors generated.

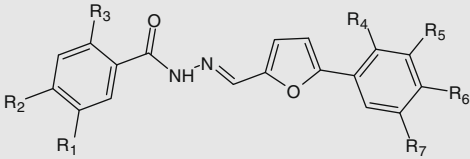
### 2.3. Linear discriminant analysis

One of the most powerful pattern recognition techniques is linear discriminant analysis (LDA), and we recently applied it to SAR analysis of compounds with elastase inhibitory activity.<sup>14</sup> Likewise, we used LDA here as a basic methodology for SAR classification of the 86 hydrazide derivatives. Taking into account that classical LDA is unable to handle as many as 836 descriptors for 86 compounds, we performed advanced LDA with the Forward Stepwise option available in STATISTICA 6.0. At each step, descriptors were successively included or excluded until no significant ( $p < 0.05$ ) improvement of the model was achieved. This procedure led to the selection of only 14 significant variables from the initially generated 836 descriptors. The following atom pairs were selected by Forward Stepwise LDA: C4\_2\_NA, C3\_3\_CL, C4\_3\_O2, CO\_3\_NA, CO\_3\_NO, C3\_5\_C4, C4\_7\_OF, BR\_11\_CA, CA\_12\_CL, CL\_12\_NA, BR\_13\_C4, C4\_14\_NO, BR\_15\_C4, and CL\_15\_O2. Use of the classification

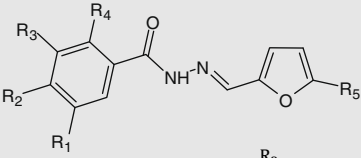
**Table 1**  
Effect of arylcarboxylic acid hydrazide derivatives on macrophage TNF- $\alpha$  production



Compound	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	R <sub>6</sub>	FI <sup>a</sup>
<i>(A) (2-Furyl)methylene-hydrazides of nicotinic acid</i>							
1	H	H	H	Br	H	H	50
2	H	H	H	Cl	CH <sub>3</sub>	H	60
3	H	Br	H	H	Cl	H	25
4	CH <sub>3</sub>	H	H	Cl	CH <sub>3</sub>	H	21
5	H	H	H	Cl	H	H	17
6	H	H	H	H	Cl	H	15
7	H	Br	H	Cl	CH <sub>3</sub>	H	10
8	CH <sub>3</sub>	H	H	CF <sub>3</sub>	H	H	<5
9	CH <sub>3</sub>	H	Cl	H	Cl	Cl	<5
10	H	H	H	COOH	OH	H	NA
11	H	H	H	H	Br	H	NA
12	CH <sub>3</sub>	H	H	H	Cl	H	NA
13	CH <sub>3</sub>	H	H	Cl	H	H	NA
14	H	H	Cl	H	Cl	Cl	NA
15	H	Br	Cl	H	Cl	H	NA
16	H	H	Cl	Cl	H	H	NA
17	H	H	Cl	H	H	H	NA
18	CH <sub>3</sub>	H	H	Br	H	H	NA
19	CH <sub>3</sub>	H	Cl	Cl	H	H	NA
20	CH <sub>3</sub>	H	Cl	H	Cl	H	NA
21	CH <sub>3</sub>	H	Cl	H	H	Cl	NA
22	CH <sub>3</sub>	H	H	Cl	Cl	H	NA

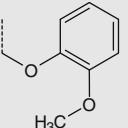


Compound	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	R <sub>6</sub>	R <sub>7</sub>	FI
<i>(B) (2-Furyl)methylene-hydrazides of benzoic acid</i>								
23	H	F	H	H	CF <sub>3</sub>	H	H	35
24	H	H	H	Cl	Cl	H	H	8
25	NO <sub>2</sub>	H	H	H	CF <sub>3</sub>	H	H	<5
26	H	Cl	Cl	Cl	H	H	H	<5
27	H	NO <sub>2</sub>	H	Cl	H	H	Cl	<5
28	I	H	Cl	H	CF <sub>3</sub>	H	H	<5
29	H	Br	H	H	CF <sub>3</sub>	H	H	NA
30	H	H	NO <sub>2</sub>	Cl	H	Cl	H	NA
31	H	H	OH	H	Cl	Cl	H	NA
32	H	NO <sub>2</sub>	H	H	Cl	H	H	NA
33	H	OCH <sub>3</sub>	H	H	COOH	Cl	H	NA
34	H	NO <sub>2</sub>	H	H	Cl	OCH <sub>3</sub>	H	NA
35	OH	H	H	H	H	NO <sub>2</sub>	H	NA
36	H	<i>t</i> -butyl	H	H	H	NO <sub>2</sub>	H	NA



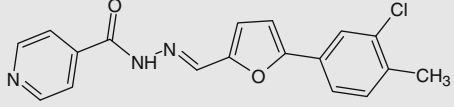
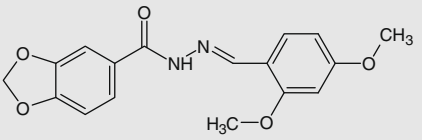
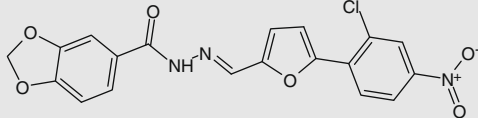
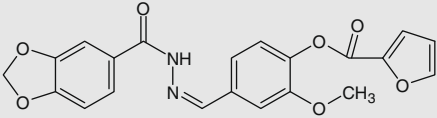
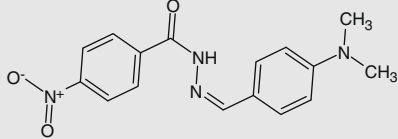
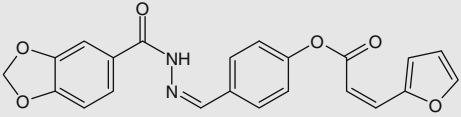
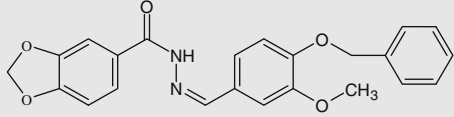
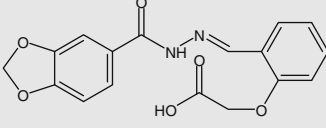
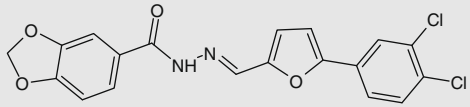
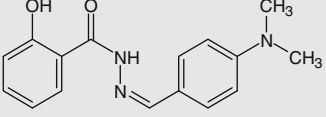
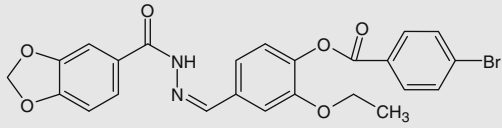
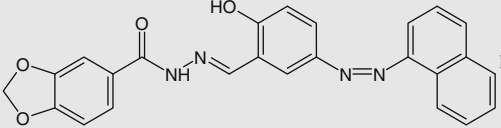
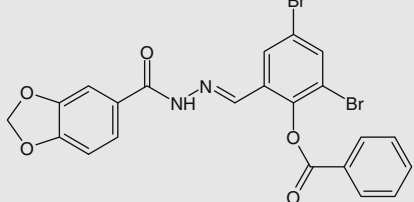
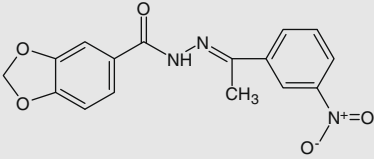
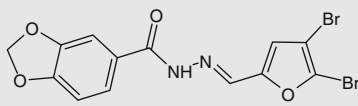
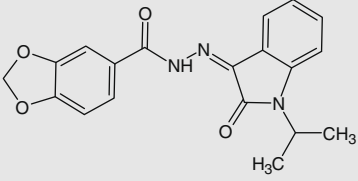
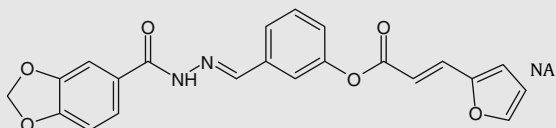
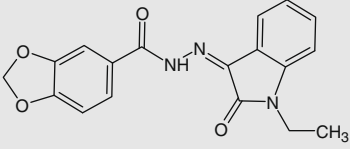
Compound	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	FI
37	H	Br	H	OH	H	NA
38	H	CH <sub>3</sub>	H	H	CH <sub>3</sub>	NA
39	H	H	H	H	H	NA
40	H	OH	H	OH	NO <sub>2</sub>	NA
41	NO <sub>2</sub>	H	NO <sub>2</sub>	OH	NO <sub>2</sub>	NA
42	H	Cl	H	H	H	NA
43	Br	H	H	OCH <sub>3</sub>	CH <sub>3</sub>	NA
44	H	Br	H	OCH <sub>3</sub>	H	NA
45	H		H	H	H	NA

Table 1 (continued)

Compound	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	FI
46	H	H	H	H	Br	NA
47	H	H	F	H	H	NA
48	H		H	H	H	NA
49	H	Cl	H	Cl	H	NA
50	H	H	H	F	I	NA

Compound	Structure	FI	Compound Structure	FI
----------	-----------	----	--------------------	----

## (C) Other derivatives

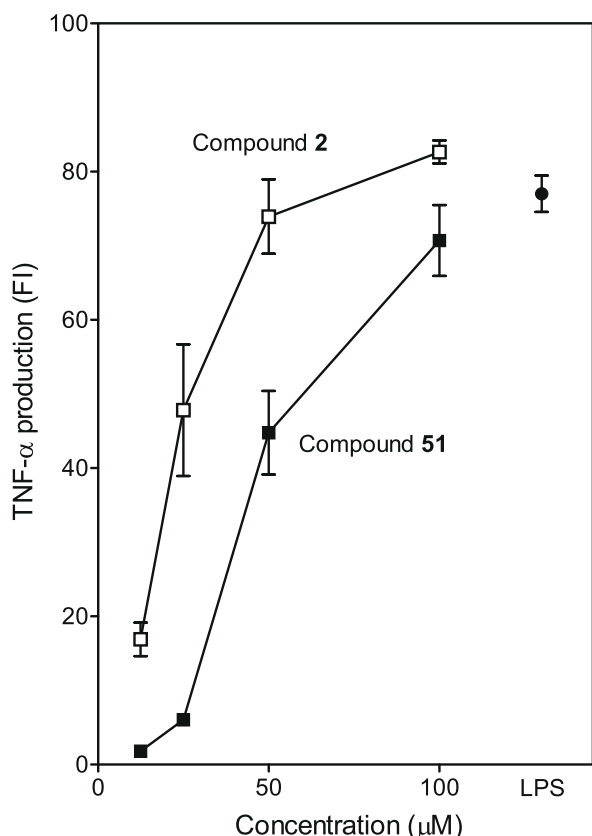
51		50	69		NA
52		35	70		NA
53		25	71		NA
54		7	72		NA
55		<5	73		NA
56		<5	74		NA
57		<5	75		NA
58		<5	76		NA
59		NA	77		NA

(continued on next page)

Table 1 (continued)

Compound	Structure	FI	Compound	Structure	FI
60		NA	78		NA
61		NA	79		NA
62		NA	80		NA
63		NA	81		NA
64		NA	82		NA
65		NA	83		NA
66		NA	84		NA
67		NA	85		NA
68		NA	86		NA

<sup>a</sup> Macrophage TNF- $\alpha$  production induced by 50  $\mu$ M of the indicated compound is shown as fold-increase (FI) above response to vehicle (DMSO) control. Activity of compounds 2–7, 11–22, 51, and 83–86 was evaluated in the present work. Data for compounds 1, 8–10, 23–50, and 52–82 were from our previous report.<sup>8</sup>



**Figure 1.** Effect of the most potent arylcarboxylic acid hydrazides on macrophage TNF- $\alpha$  production. J774.A1 macrophages ( $2 \times 10^5$  cells/well) were cultured in the presence of the indicated concentrations of compound **2** ( $\square$ ), compound **51** ( $\blacksquare$ ), or 50 ng/ml LPS ( $\bullet$ ) for 24 h, and TNF- $\alpha$  was measured in the cell supernatants by ELISA. The data are presented as the fold-increase (FI) in TNF- $\alpha$  production above DMSO control and represent means  $\pm$  SD of three independent experiments with triplicate samples analyzed in each experiment.

functions obtained with these pairs resulted in 95.3% correct classification: 20 of 23 active and 62 of 63 inactive hydrazides were correctly classified to their experimentally determined activity. In addition, values of the 14 atom pairs selected were not mutually correlated with each other ( $r \leq 0.7$ ), that is, they can be regarded as independent variables.

In order to further decrease the number of descriptors, we performed LDA analysis with the Best Subset Search option, starting from 14 atom pairs selected after the first run of the LDA procedure and found that the best subset consisted of 13 atom pairs as listed above, but with C4\_2\_NA excluded. These variables provided the least misclassification error among all other possible subsets of different sizes chosen from 14 descriptors, and the SAR model obtained had an improved quality of classification: 96.5% compounds were classified correctly compared to their experimental activity (Tables 2 and 3). This LDA model can be presented by two classification functions  $F(\text{Active})$  and  $F(\text{NA})$ :

$$\begin{aligned}
 F(\text{Active}) = & -11.81 + 15.52\text{C3\_3\_CL} - 16.49\text{C4\_3\_O2} \\
 & + 5.62\text{CO\_3\_NA} - 19.42\text{CO\_3\_NO} \\
 & + 19.08\text{C3\_5\_C4} - 9.11\text{C4\_7\_OF} \\
 & + 10.82\text{BR\_11\_CA} + 2.17\text{CA\_12\_CL} \\
 & - 22.61\text{CL\_12\_NA} - 9.26\text{BR\_13\_C4} \\
 & + 10.96\text{C4\_14\_NO} - 15.18\text{BR\_15\_C4} \\
 & + 7.34\text{CL\_15\_O2}
 \end{aligned} \quad (1)$$

$$\begin{aligned}
 F(\text{NA}) = & -0.725 + 0.567\text{C3\_3\_CL} + 0.514\text{C4\_3\_O2} \\
 & + 1.213\text{CO\_3\_NA} + 0.160\text{CO\_3\_NO} + 1.648\text{C3\_5\_C4} \\
 & + 0.859\text{C4\_7\_OF} + 1.237\text{BR\_11\_CA} + 0.478\text{CA\_12\_CL} \\
 & - 0.915\text{CL\_12\_NA} + 0.060\text{BR\_13\_C4} \\
 & + 1.229\text{C4\_14\_NO} - 0.744\text{BR\_15\_C4} \\
 & + 0.802\text{CL\_15\_O2}
 \end{aligned} \quad (2)$$

According to these equations, a compound will be classified as 'Active' if the value of  $F(\text{Active}) > F(\text{NA})$ , and vice versa. The classifications observed and calculated by the LDA model for compounds **1–86** are shown in Table 3, and values of all atom pair descriptors used in Eqs. 1 and 2 are shown in Supplementary Table S1.

The predictive ability of the LDA model was evaluated by the leave-one-out (LOO) procedure. The LOO prediction resulted in 89.5% correct classification, and 18 of 23 active and 59 of 63 inactive hydrazides were correctly predicted for their TNF- $\alpha$  induction activity classes (Table 3). Thus, these results confirm usefulness of the LDA model for a priori evaluation of macrophage TNF- $\alpha$  inducing activity of arylcarboxylic acid hydrazides.

Although 13 atom pair descriptors were utilized in the derived LDA model, this number should not be regarded as too large. Conventionally, the recommended number of variables for SAR and QSAR models, from a statistical point of view, should be  $\leq 20\%$  of the number of compounds. Hence, the number of atom pairs selected is reasonable for 86 hydrazide derivatives investigated. Additionally, all coefficients of the classification functions (Eqs. 1 and 2) were significant according to the Fisher criterion.

The atom pairs involved in Eqs. 1 and 2 are not uniformly distributed in the number of chemical bonds  $D$ . Figure 3B shows that six atom pairs used in the LDA model have bond distances from 3 to 7, while the other seven descriptors are characterized by  $D$  values from 11 to 15. Indeed, this distribution is a reflection of total atom pair distribution (Fig. 3A), which is conditioned by the dumb-bell shape of the compounds investigated. On the other hand, the importance of 'longer' atom pairs for SAR classification supports the supposition that a biological target interacts with the entire hydrazide molecule, rather than with metabolites of a smaller size.

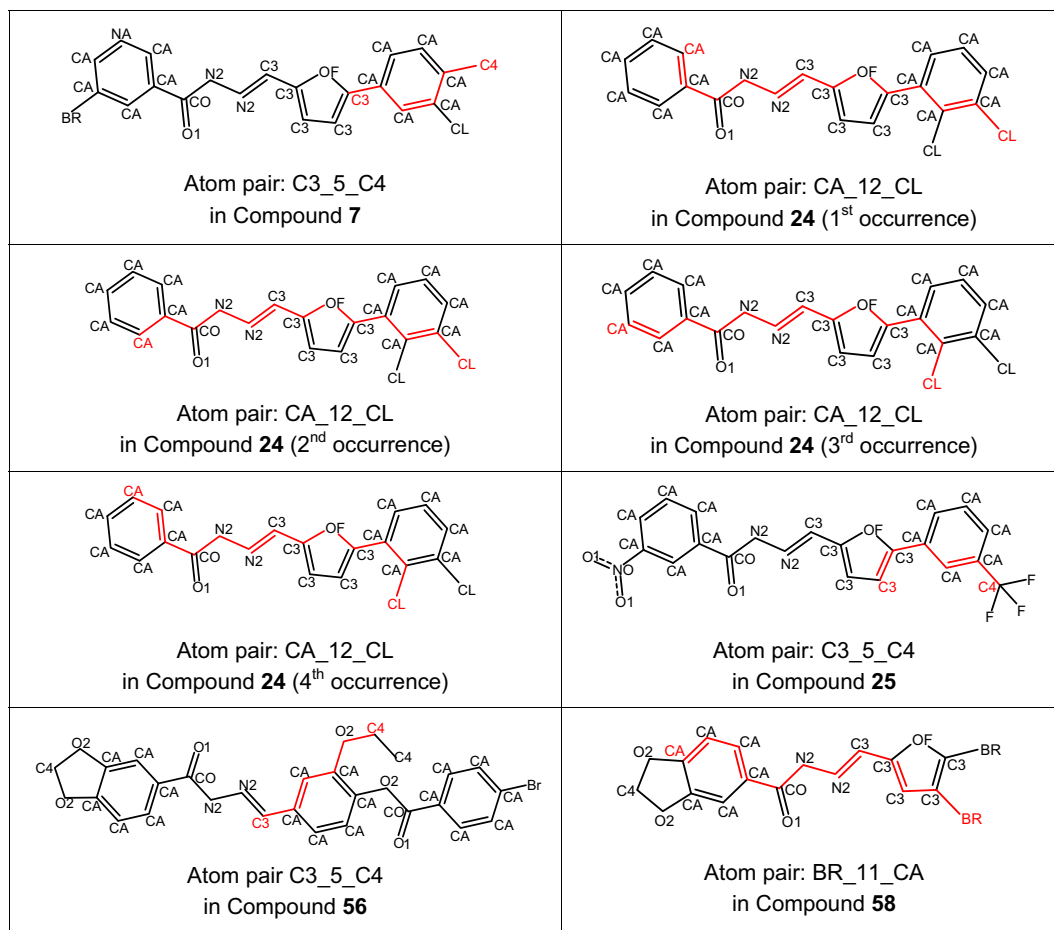
## 2.4. Classification tree analysis with linear combination splits

In our previous SAR analysis of *N*-benzoylpyrazoles with elastase inhibitory activity, we also used LDA methodology,<sup>14</sup> however, its use was preceded by application of one-way analysis of variance (ANOVA)<sup>16</sup> for preliminary selection of descriptors having significant differences between in-class and total variances. This led to a substantial decrease in the number of atom pairs to reduce dimensionality of the data matrix for further SAR analysis. Since each descriptor selected by ANOVA has one-dimensional separation of classes, compounds from different groups are characterized by relatively distinct areas of data point projections on a single coordinate axis associated with a given descriptor (e.g., see Fig. 4A).

It should be noted that in the case of hydrazides **1–86**, the pre-selection of atom pairs by ANOVA did not result in a satisfactory SAR model for predicting their macrophage TNF- $\alpha$  inducing activity if the LDA method was applied to the ANOVA-selected descriptors. Instead, good classification was achieved by stepwise LDA applied to the initial non-reduced data matrix, as described above. Notably, only 3 atom pairs (C3\_5\_C4, CA\_12\_CL, and C4\_14\_NO) of 13 descriptors involved in Eqs. 1 and 2 were selected in the trial run of ANOVA and thus had approximately one-dimensional class separation, as exemplified in Figure 4A.

The other 10 atom pairs had occurrences that non-significantly differed between classes of active and non-active compounds. These atom pair descriptors clearly belong to another type where the





**Figure 2.** Examples of atom pair descriptors in selected active arylcarboxylic acid hydrazides. Atom pairs are depicted in red and indicated below the structure. Compound numbers correspond to those shown in Table 1.

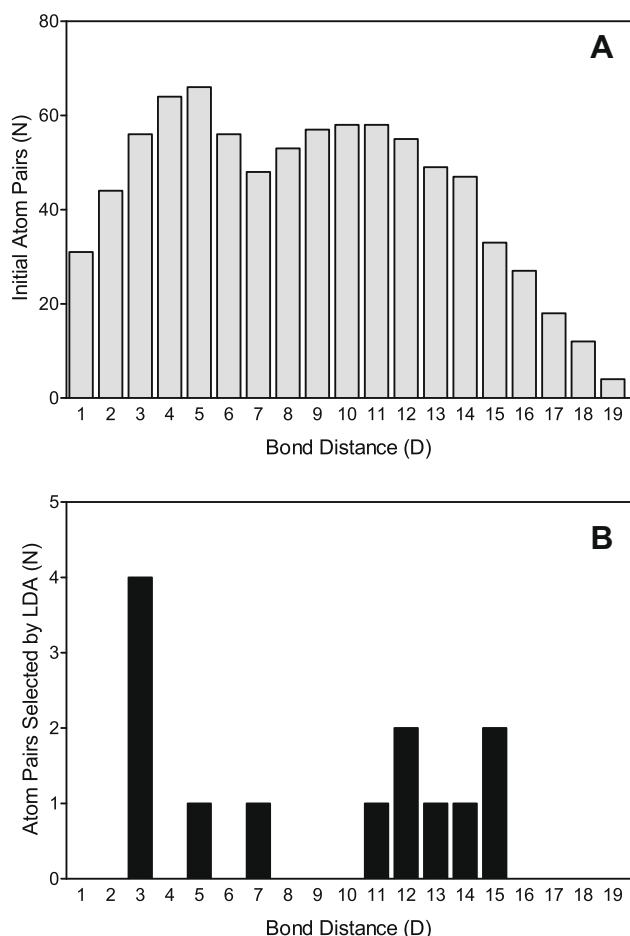
activity classes were separated in higher-dimensional subspaces of such descriptors (see two-dimensional example in Fig. 4B). Although projections of data points for both classes in this example are approximately uniformly distributed on each coordinate axis, there exists a line of good separation, and such descriptors appear to be very useful for SAR analysis, as demonstrated above by LDA. In a more common case of higher dimensionality, there may exist a hyper-plane separating two classes of compounds. Taking into account the distribution character of data points for compounds 1–86 in descriptor space, we attempted to apply a methodology known as classification tree analysis with linear combination splits (CTLCS).<sup>17</sup>

In this approach, a logical tree was created where a split condition for each tree node depends on a linear combination of several descriptors. We found that the best classification tree for compounds 1–86 had just one split. The 13 atom pairs utilized in the LDA model (see Eqs. 1 and 2) were used as a basis in the CTLCS approach, and all pairs were included in the function  $F(x)$  (Eq. 3), indicating again that all 13 descriptors were important for prediction of the correct biological activity class.

$$\begin{aligned}
 F(x) = & -0.122 + 0.192C3\_3\_CL - 0.219C4\_3\_O2 \\
 & + 0.057CO\_3\_NA - 0.252CO\_3\_NO + 0.224C3\_5\_C4 \\
 & - 0.128C4\_7\_OF + 0.123BR\_11\_CA + 0.022CA\_12\_CL \\
 & - 0.279CL\_12\_NA - 0.120BR\_13\_C4 + 0.125C4\_14\_NO \\
 & - 0.186BR\_15\_C4 + 0.084CL\_15\_O2
 \end{aligned} \quad (3)$$

According to the split condition, a compound would be classified as inactive if  $F(x) \leq 0$ ; otherwise a compound belongs to the 'Active' class. The classification matrix obtained by the CTLCS method is shown in Table 2. The activity classes were predicted correctly for 20 of 23 active and 59 of 63 inactive hydrazides, resulting in a total accuracy of fitting 91.9%. The calculated and LOO-predicted classes for individual compounds are shown in Table 3. In 73 of 86 cases (84.9%), a priori prediction of activity class by the LOO procedure was correct. While LDA classification by Eqs. 1 and 2 had better characteristics of fitting and prediction (Table 2), the CTLCS model was twofold simpler in the amount of calculation necessary for a compound classification. Satisfactory results obtained by the one-split tree based on linear combination of variables indicates that the descriptor space is divided into two areas by a hyper-plane expressed by Eq. 3. Each of these areas preferentially contains data points for compounds of a single activity class, such as in the simulated two-dimensional example given in Figure 4B. Such well-organized data in a space of atom pair descriptors demonstrates the powerful ability of atom pairs to separate compounds of different activity in SAR analysis.

It should be noted that most of the incorrect classifications by both the LDA and CTLCS methods were made in the subset of nicotinic acid hydrazide derivatives 1–22 (Table 3). Hence, some structural or physicochemical peculiarities of nicotinic acid hydrazides (e.g., polarizability, dipole moment, etc.) may be reflected non-significantly in the entire matrix of atom pair descriptors.



**Figure 3.** Numbers of unique atom pairs in the set of arylcarboxylic acid hydrazides. The numbers are shown for each of the indicated bond distances initially generated for the 86 hydrazides (A). Atom pairs subsequently included in the best LDA model are shown in (B).

## 2.5. Classification tree analysis with univariate splits

Although the LDA and CTLCS models had high fitting and predictive abilities, it is difficult to formulate these models in a set of intuitively understandable ‘chemical’ rules. The methodology of binary classification tree analysis with univariate splits<sup>18</sup> is more suitable for deriving simplified SAR rules, while being less complex than the LDA or CTLCS methods. Based on the 13 descriptors selected in LDA above, we obtained the optimal classification tree with univariate splits shown in Figure 5. The atom pair descriptors involved in the optimal tree were selected automatically by STATISTICA 6.0 using an exhaustive univariate split selection method (see Section 4).

According to this tree, the prediction of compounds **1–86** as ‘Active’ or ‘NA’ depends on three atom pairs: C3\_5\_C4, CA\_12\_CL, and

BR\_11\_CA (examples shown in Fig. 2). Taking into account that atom pair descriptors adopt integer values only, the conditions present in Figure 5 can be interpreted as follows. If a compound has at least one C3\_5\_C4 atom pair, then the compound is classified as ‘Active.’ Similarly, on the second and third splits, a compound is classified as ‘Active’ if it has more than three CA\_12\_CL atom pairs or more than one BR\_11\_CA atom pair, respectively. An insufficient number of all the enumerated atom pairs leads to the left lowest terminal node where the compound is assigned as ‘NA.’ In total, 84.9% of the compounds were classified correctly using only these three atom pairs (57 of 63 inactive and 16 of 23 active arylcarboxylic acid hydrazide derivatives were correctly classified) (Table 2). Classifications made by the tree for compounds **1–86** are shown in Table 3.

As indicated above, the BR\_11\_CA atom pair represents in a ‘chemical’ sense the simultaneous presence of a bromine atom and an aromatic ring on the opposite sides of a molecule. The descriptor C3\_5\_C4 is characteristic of two types of compounds: one containing a methyl- or trifluoromethyl-substituted benzene ring attached to the furan moiety (see Fig. 2, compounds **7** and **25**), and the other containing an alkoxy group in the aromatic ring connected to the azomethine carbon of the linker (see Fig. 2, compound **56**). If activity is based on the presence of the CA\_12\_CL atom pair, at least four of these atom pairs are necessary for classification as ‘Active.’ This atom pair is present when aromatic fragments are located on both sides of a dumbbell-shaped molecule, with one aromatic moiety containing two or more chlorine atoms in *ortho* and *meta* positions (four such atom pairs in compound **24** are shown in Fig. 2).

Although the accuracy of classification by these simplified rules is slightly lower than that of the LDA or CTLCS approaches, it can be very useful for non-computational, logical prediction of the activity class for a given arylcarboxylic acid hydrazide derivative. Note that the classification tree model with univariate splits, like the LDA and CTLCS models, also includes ‘longer’ atom pairs with 11 and 12 chemical bonds, which is in agreement with the proposed interaction of the entire non-metabolized molecule with a given biological target rather than smaller metabolites.

## 3. Conclusion

Previously, we identified a novel class of compounds that potentially induced TNF- $\alpha$  production in macrophages via activation of *N*-formyl peptide receptors and found that the active compounds had an arylcarboxylic acid hydrazide core structure.<sup>8</sup> Here, we identified additional arylcarboxylic acid hydrazide derivatives that induced macrophage TNF- $\alpha$  production. We then used the combined group of all 86 compounds for SAR analysis to further define the features of these molecules important for activity and developed a simple, but accurate SAR model for predicting biological activity in future compound screening. A sequence of LDA, classification tree analyses with linear combination, and univariate splits based on the atom pair descriptors led to the derivation of SAR rule-based algorithms with 96.5%, 91.9%, and 84.9% predictive

**Table 2**

Classification matrices for linear discriminant analysis (LDA) and classification tree analyses with linear combination splits (CTLCS) univariate splits

Experimentally determined classification	Calculated classification								
	LDA model			CTLCS model			Classification tree with univariate splits		
	Active	NA	Accuracy (%)	Active	NA	Accuracy (%)	Active	NA	Accuracy (%)
Active	<b>20</b>	3	87.0	<b>20</b>	3	87.0	<b>16</b>	7	69.6
NA	0	<b>63</b>	100.0	4	<b>59</b>	93.7	6	<b>57</b>	90.5
Total	20	66	96.5	24	62	91.9	22	64	84.9

The number of compounds correctly classified by the model is indicated in bold.



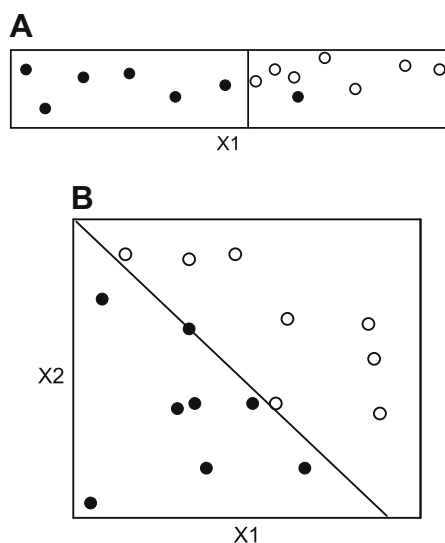
**Table 3**Experimentally determined, SAR-calculated, and LOO-predicted classes of macrophage TNF- $\alpha$  inducing activity for all 86 arylcarboxylic acid hydrazide derivatives

Compound	Determined	LDA Model		CTLCS Model		Classification Tree with Univariate Splits			
		Calculated	LOO-predicted	Calculated	LOO-predicted	Atom pairs and frequency of occurrence			Calculated <sup>a</sup>
						C3_5_C4	CA_12_CL	BR_11_CA	
1	Active	Active	Active	Active	Active	0	0	1 →	NA
2	Active	Active	Active	Active	Active	1		→	Active
3	Active	Active	Active	Active	Active	0	1	1 →	NA
4	Active	Active	Active	Active	Active	1		→	Active
5	Active	NA	NA	NA	NA	0	2	0 →	NA
6	Active	NA	NA	NA	NA	0	1	0 →	NA
7	Active	Active	Active	Active	Active	1		→	Active
8	Active	Active	Active	Active	Active	1		→	Active
9	Active	NA	NA	NA	NA	0	4	→	Active
10	NA	NA	NA	NA	NA	0	0	0 →	NA
11	NA	NA	NA	NA	NA	0	0	0 →	NA
12	NA	NA	NA	NA	NA	0	1	0 →	NA
13	NA	NA	NA	NA	NA	0	2	0 →	NA
14	NA	NA	NA	NA	NA	0	4	→	Active
15	NA	NA	Active	Active	Active	0	2	1 →	NA
16	NA	NA	NA	NA	NA	0	3	0 →	NA
17	NA	NA	NA	NA	NA	0	1	0 →	NA
18	NA	NA	NA	NA	NA	0	0	1 →	NA
19	NA	NA	NA	NA	NA	0	3	0 →	NA
20	NA	NA	NA	NA	NA	0	2	0 →	NA
21	NA	NA	NA	NA	NA	0	3	0 →	NA
22	NA	NA	NA	Active	Active	0	3	0 →	NA
23	Active	Active	Active	Active	Active	1		→	Active
24	Active	Active	Active	Active	Active	0	4	→	Active
25	Active	Active	Active	Active	Active	1		→	Active
26	Active	Active	Active	Active	Active	0	5	→	Active
27	Active	Active	Active	Active	Active	0	4	→	Active
28	Active	Active	Active	Active	Active	1		→	Active
29	NA	NA	NA	NA	NA	1		→	Active
30	NA	NA	Active	NA	Active	0	3	0 →	NA
31	NA	NA	NA	NA	NA	0	3	0 →	NA
32	NA	NA	NA	NA	NA	0	2	0 →	NA
33	NA	NA	NA	NA	NA	0	1	0 →	NA
34	NA	NA	NA	NA	NA	0	2	0 →	NA
35	NA	NA	NA	NA	NA	0	0	0 →	NA
36	NA	NA	NA	NA	NA	0	0	0 →	NA
37	NA	NA	NA	NA	NA	0	0	0 →	NA
38	NA	NA	NA	NA	NA	0	0	0 →	NA
39	NA	NA	NA	NA	NA	0	0	0 →	NA
40	NA	NA	NA	NA	NA	0	0	0 →	NA
41	NA	NA	NA	NA	NA	0	0	0 →	NA
42	NA	NA	NA	NA	NA	0	0	0 →	NA
43	NA	NA	NA	NA	NA	0	0	0 →	NA
44	NA	NA	NA	NA	NA	0	0	0 →	NA
45	NA	NA	NA	NA	NA	0	0	0 →	NA
46	NA	NA	NA	Active	Active	0	0	1 →	NA
47	NA	NA	NA	NA	NA	0	0	0 →	NA
48	NA	NA	NA	NA	NA	0	0	0 →	NA
49	NA	NA	NA	NA	NA	0	0	0 →	NA
50	NA	NA	NA	NA	NA	0	0	0 →	NA
51	Active	Active	Active	Active	Active	1		→	Active
52	Active	Active	Active	Active	Active	0	2	0 →	NA
53	Active	Active	NA	Active	NA	0	0	0 →	NA
54	Active	Active	Active	Active	Active	1		→	Active
55	Active	Active	NA	Active	NA	0	3	0 →	NA
56	Active	Active	Active	Active	Active	1		→	Active
57	Active	Active	Active	Active	Active	0	0	2 →	Active
58	Active	Active	Active	Active	Active	0	0	2 →	Active
59	NA	NA	NA	NA	NA	0	0	0 →	NA
60	NA	NA	Active	NA	Active	1		→	Active
61	NA	NA	NA	NA	NA	0	0	0 →	NA
62	NA	NA	NA	NA	NA	0	0	1 →	NA
63	NA	NA	NA	NA	NA	0	0	0 →	NA
64	NA	NA	NA	NA	NA	0	0	0 →	NA
65	NA	NA	NA	NA	NA	0	0	2 →	Active
66	NA	NA	NA	NA	NA	0	0	0 →	NA
67	NA	NA	NA	NA	NA	0	0	0 →	NA
68	NA	NA	NA	NA	NA	0	0	0 →	NA
69	NA	NA	NA	NA	NA	0	0	0 →	NA
70	NA	NA	Active	NA	Active	1		→	Active
71	NA	NA	NA	NA	NA	0	0	0 →	NA
72	NA	NA	NA	NA	NA	0	0	0 →	NA

Table 3 (continued)

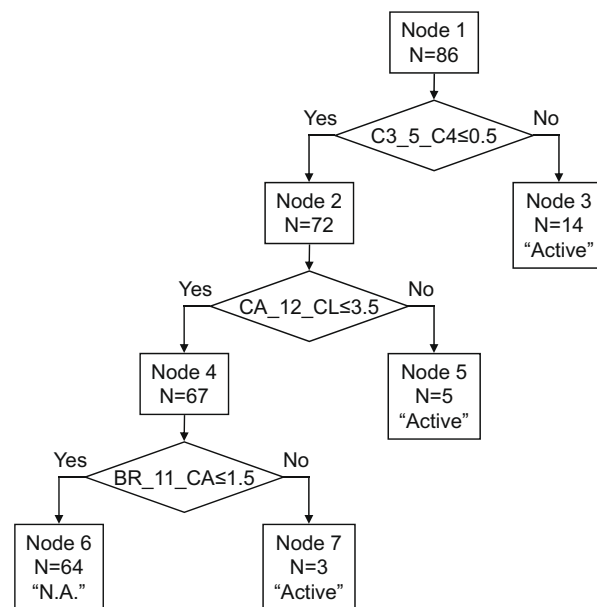
Compound	Determined	LDA Model		CTLCS Model		Classification Tree with Univariate Splits			
		Calculated	LOO-predicted	Calculated	LOO-predicted	Atom pairs and frequency of occurrence			Calculated <sup>a</sup>
						C3_5_C4	CA_12_CL	BR_11_CA	
73	NA	NA	NA	NA	NA	0	0	0 →	NA
74	NA	NA	NA	NA	NA	0	0	0 →	NA
75	NA	NA	NA	NA	NA	0	0	0 →	NA
76	NA	NA	NA	NA	NA	0	0	0 →	NA
77	NA	NA	NA	NA	NA	0	0	0 →	NA
78	NA	NA	NA	NA	NA	0	0	0 →	NA
79	NA	NA	NA	NA	NA	0	0	0 →	NA
80	NA	NA	NA	NA	NA	0	0	0 →	NA
81	NA	NA	NA	NA	NA	0	0	0 →	NA
82	NA	NA	NA	NA	Active	0	1	0 →	NA
83	NA	NA	NA	NA	NA	0	1	0 →	NA
84	NA	NA	NA	NA	NA	0	3	0 →	NA
85	NA	NA	NA	NA	NA	1			Active
86	NA	NA	NA	Active	Active	0	0	1 →	NA

<sup>a</sup> Incorrect classifications are indicated in italics. Arrows correspond to compound classification upon entering terminal nodes of the tree shown in Figure 5.



**Figure 4.** Simulated examples of descriptors with one-dimensional (A) and two-dimensional (B) separation. Active and non-active compounds are represented by open and close circles, respectively.

accuracy, respectively. Furthermore, LOO analysis confirmed the usefulness of these models for a priori evaluation of macrophage TNF- $\alpha$  inducing activity by arylcarboxylic acid hydrazides. The intuitively understandable rules obtained from the classification tree with univariate splits, which is based on three atom pair descriptors only, revealed that the main factors influencing the activity of a given arylcarboxylic acid hydrazide derivative were either (1) the presence of a methyl or trifluoromethyl group in the benzene ring attached to the furan moiety, (2) an alkoxy group in the aromatic ring near the methylenehydrazide linker, or (3) two or more halogen atoms (chlorine or bromine) in one side of the dumbbell-shaped molecule, with an aromatic fragment on the opposite side. The successful application of atom pairs to heterogeneous sets of compounds can be explained by their non-global nature, as this approach is based on simple local features of molecules rather than on certain chemical building blocks. Overall, our data demonstrate that the use of atom pair descriptors is a valuable tool for developing different SAR rules for high-throughput screening of data sets and could provide a relatively simple classification useful for de novo design of macrophage TNF- $\alpha$  inducers with arylcarboxylic acid hydrazide scaffolds.



**Figure 5.** Binary classification tree reflecting the simplified SAR rules for predicting macrophage TNF- $\alpha$  inducing activity of arylcarboxylic acid hydrazide derivatives.

## 4. Experimental

### 4.1. Reagents

The additional 23 compounds (**2–7**, **11–22**, **51**, and **83–86**) investigated were purchased from Princeton BioMolecular Research, Inc. (Monmouth Junction, NJ). Their purity and identity were verified by Princeton BioMolecular Research using NMR spectroscopy, elemental analysis, and mass spectroscopy. <sup>1</sup>H NMR spectra provided by Princeton BioMolecular Research for these compounds (10% solutions in deuterated dimethyl sulfoxide, DMSO-*d*<sub>6</sub>) were obtained with a Bruker Avance 200 MHz spectrometer (Bruker BioSpin, Billerica, MA) and are included in [Supplementary Table S2](#).

### 4.2. Cell culture

Murine macrophage J774.A1 cells were cultured in DMEM supplemented with 10% (v/v) heat-inactivated fetal bovine serum (FBS), 10 mM HEPES, 100  $\mu$ g/ml streptomycin, and 100 U/ml

penicillin. Cells were grown in sterile tissue culture flasks at 37 °C in a humidified atmosphere containing 5% CO<sub>2</sub> and gently detached by scraping.

#### 4.3. Determination of TNF- $\alpha$

For treatments, cells were plated in 96-well microtiter plates at  $2 \times 10^5$  cells/well in culture media, except FBS was reduced to 3% (v/v). The cells were treated for 24 h with negative control DMSO, test compounds, or positive control LPS. A murine TNF- $\alpha$  enzyme-linked immunosorbent assay (ELISA) kit (BD Biosciences Pharmingen) was used to detect this cytokine in the cell supernatants. Cytokine concentrations were determined by extrapolation from the TNF- $\alpha$  standard curve, according to the manufacturer's protocol.

#### 4.4. Cytotoxicity assay

Cytotoxicity was analyzed with a CellTiter-Glo Luminescent Cell Viability Assay Kit (Promega, Inc., Madison, WI), according to the manufacturer's protocol. Briefly, J774.A1 cells were cultured at a density of  $3 \times 10^4$  cells/well with the test compounds for 24 h at 37 °C and 5% CO<sub>2</sub>, substrate was added, and luminescence signal in the samples was analyzed with a Fluoroscan Ascent FL microplate reader.

#### 4.5. Endotoxin assay

Endotoxin was measured using the Limulus Amebocyte Lysate Pyrogen Plus kit (Cambrex Bio Science, Walkersville, MD). Briefly, the limulus amebocyte lysate was reconstituted in 250  $\mu$ l solution of test compound (50  $\mu$ M in endotoxin-free water/1% DMSO), and each vial was incubated at 37 °C for 1 h. At the end of the incubation period, each vial was inverted 180° to estimate gel formation in comparison with control (endotoxin-free water).

#### 4.6. Structure encoding by atom pairs

For the purpose of SAR analysis, we used an atom pair representation of molecular structures with each atom pair denoted as T1\_D\_T2, where T1 and T2 are the types of atoms in the pair and D is the topological (bond) distance or number of bonds in the shortest path between these atoms in the structural formula. As previously reported,<sup>14</sup> T1 and T2 were defined with symbolic codes used in HyperChem, Version 7 (Hypercube, Inc., Gainesville, FL) for atom type representation within MM+ force field. For example, CA, CO, and C3 codes were used for sp<sup>2</sup>-hybridized aromatic, carbonyl, and furan carbon atoms, respectively. This approach allows easy generation of atom pairs directly from the output file containing the molecular structure (HIN file) built by HyperChem. As atom pairs T1\_D\_T2 and T2\_D\_T1 are equivalent, we used a unified definition with lexicographic order of type substrings (i.e., with T1  $\leq$  T2).

All 836 unique atom pairs possible for non-hydrogen atoms in the 86 derivatives of arylcarboxylic acid hydrazides were generated. This  $86 \times 836$  data matrix was automatically built by our CHAIN program, based on HIN files created in HyperChem. A matrix element at the intersection of the *i*th row and *j*th column was equal to the *j*th atom pair occurrence in the *i*th molecule.

#### 4.7. Derivation of SAR classification

Derivation of SAR classification was performed first by the LDA method with the 'Forward Stepwise' option, using the corresponding module of STATISTICA 6.0. The statistical criterion for inclusion or exclusion of descriptors at each step was  $p \leq 0.05$ . The stepwise LDA allowed selection of 14 significant descriptors from 836 atom pairs generated initially. The LDA run was then repeated with the 'Best Subset Search' option on the basis of 14 variables selected in the first LDA run. The best subset consisted of 13 atom pairs giving the least misclassification error of LDA model.

Starting from 13 variables of the best subset, we developed binary classification tree models with discriminant-based linear combination splits (CTLCS) and with univariate splits. The classification trees were built with STATISTICA 6.0 using estimated prior probabilities and equal misclassification costs for classes.<sup>17,18</sup> An exhaustive C&RT-style univariate split selection method was used, as described by Breiman et al.<sup>18</sup>

#### Acknowledgments

This work was supported in part by Department of Defense Grant W9113M-04-1-0001, National Institutes of Health Grants P20 RR-020185 and U54 AI-065357, National Institutes of Health contract HHSN266200400009C, an equipment grant from the M.J. Murdock Charitable Trust, and the Montana State University Agricultural Experimental Station. The U.S. Army Space and Missile Defense Command, 64 Thomas Drive, Frederick, MD 21702 is the awarding and administering acquisition office. The content of this report does not necessarily reflect the position or policy of the U.S. Government.

#### Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2008.08.078.

#### References and notes

- Beutler, B. J. *Invest. Med.* **1995**, 43, 227.
- Wagner, G.; Laufer, S. *Med. Res. Rev.* **2006**, 26, 1.
- Lejeune, F. J.; Lienard, D.; Matter, M.; Ruegg, C. *Cancer Immun.* **2006**, 6, 6.
- Reed, J. C. *Nat. Clin. Pract. Oncol.* **2006**, 3, 388.
- Baguley, B. C. *Curr. Opin. Invest. Drugs* **2001**, 2, 967.
- Burkhart, C. A.; Berman, J. W.; Swindell, C. S.; Horwitz, S. B. *Cancer Res.* **1994**, 54, 5779.
- Schön, M.; Schön, M. P. *Curr. Med. Chem.* **2007**, 14, 681.
- Schepetkin, I. A.; Kirpotina, L. N.; Tian, J.; Khlebnikov, A. I.; Ye, R. D.; Quinn, M. T. *Mol. Pharm.* **2008**, 74, 392.
- Andricopulo, A. D.; Montanari, C. A. *Mini. Rev. Med. Chem.* **2005**, 5, 585.
- Kirikae, T.; Ojima, I.; Kirikae, F.; Ma, Z.; Kuduk, S. D.; Slater, J. C.; Takeuchi, C. S.; Bounaud, P. Y.; Nakano, M. *Biochem. Biophys. Res. Commun.* **1996**, 227, 227.
- Ojima, I.; Fumero-Oderda, C. L.; Kuduk, S. D.; Ma, Z.; Kirikae, F.; Kirikae, T. *Bioorg. Med. Chem.* **2003**, 11, 2867.
- Buttingsrud, B.; Ryeng, E.; King, R. D.; Alsberg, B. K. *J. Comput. Aided Mol. Des.* **2006**, 20, 361.
- Khlebnikov, A. I.; Schepetkin, I. A.; Domina, N. G.; Kirpotina, L. N.; Quinn, M. T. *Bioorg. Med. Chem.* **2007**, 15, 1749.
- Khlebnikov, A. I.; Schepetkin, I. A.; Quinn, M. T. *Bioorg. Med. Chem.* **2008**, 16, 2791.
- Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. J. *Chem. Inf. Comput. Sci.* **1985**, 25, 64.
- Lindman, H. R. *Analysis of Variance in Complex Experimental Designs*; W.H. Freeman & Co.: San Francisco, 1974.
- Loh, W. Y.; Shih, Y. S. *Stat. Sin.* **1997**, 7, 815.
- Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth & Brooks/Cole Advanced Books & Software: Monterey, 1984.