

# Structure–activity relationship analysis of *N*-benzoylpyrazoles for elastase inhibitory activity: A simplified approach using atom pair descriptors

Andrei I. Khlebnikov,<sup>a,\*</sup> Igor A. Schepetkin<sup>b</sup> and 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

Received 24 October 2007; revised 7 January 2008; accepted 7 January 2008

Available online 15 January 2008

**Abstract**—Previously, we utilized high throughput screening of a chemical diversity library to identify potent inhibitors of human neutrophil elastase and found that many of these compounds had *N*-benzoylpyrazole core structures. We also found individual ring substituents had significant impact on elastase inhibitory activity and compound stability. In the present study, we utilized computational structure–activity relationship (SAR) analysis of a series of 53 *N*-benzoylpyrazole derivatives to further optimize these lead molecules. We present an improved approach to SAR methodology based on atom pair descriptors in combination with 2-dimensional (2D) molecular descriptors. This approach utilizes the rich representation of chemical structure and leads to SAR analysis that is both accurate and intuitively easy to understand. A sequence of ANOVA, linear discriminant, and binary classification tree analyses of the molecular descriptors led to the derivation of SAR rule-based algorithms. These rules revealed that the main factors influencing elastase inhibitory activity of *N*-benzoylpyrazole molecules were the presence of methyl groups in the pyrazole moiety and *ortho*-substituents in the benzoyl radical. Furthermore, our data showed that physicochemical characteristics (energy of frontier molecular orbitals, molar refraction, lipophilicity) were not necessary for achieving good SAR, as comparable quality of SAR classification was obtained with atom pairs and 2D descriptors only. This simplified SAR approach may be useful to qualitative SAR recognition problems in a variety of data sets.

© 2008 Elsevier Ltd. All rights reserved.

## 1. Introduction

Neutrophil elastase (EC 3.4.21.37) is a member of the chymotrypsin family of serine proteases, which can degrade a variety of extracellular matrix proteins and proteolytically activate several matrix metalloproteinases (MMP-2, -3, and -9) (reviewed in Refs. 1,2). Excessive neutrophil elastase activity can lead to severe pathology through the degradation of elastin and collagen in the airways, resulting in microvascular injury and interstitial edema.<sup>3</sup> Given the destructive potential of unregulated neutrophil elastase, it is not surprising that inhibition of the elastase activity in pulmonary tissues has been considered a promising strategy to improve the outcome

of pulmonary diseases.<sup>4</sup> For example, many types of peptide and nonpeptide inhibitors of neutrophil elastase, employing both reversible and irreversible mechanisms of action, have been identified (reviewed in Refs. 5,6). Recently, we utilized high throughput screening of a chemical diversity library containing 10,000 drug-like molecules and identified 19 potent neutrophil elastase inhibitors ( $K_i \leq 1 \mu\text{M}$ ) that have *N*-benzoylpyrazole core structures and are distinct from currently known elastase inhibitors.<sup>7</sup>

Our analysis of *N*-benzoylpyrazole derivatives showed that individual ring substituents had significant impact on elastase inhibitory activity and compound stability.<sup>7</sup> Thus, we suggest that further structure–activity relationship (SAR) analysis of *N*-benzoylpyrazoles would lead to optimization of these lead compounds to identify improved neutrophil elastase inhibitors. Indeed, SAR and quantitative SAR (QSAR) models have been instrumental in understanding the molecular mechanism of action of receptor antagonists, their design, and virtual screen-

**Keywords:** Atom pairs; Molecular descriptors; Structure–activity relationship; *N*-Benzoylpyrazoles; Neutrophil elastase inhibitors.

\* Corresponding authors. Tel.: +7 3852 245513/522436; fax: +7 3852 367864 (A.I.K.); Tel.: +1 406 994 5721; fax: +1 406 994 4303 (M.T.Q.); e-mail addresses: [aikhl@chem.org.ru](mailto:aikhl@chem.org.ru); [mquinn@montana.edu](mailto:mquinn@montana.edu)

ing.<sup>8</sup> (Q)SAR refers to a broad range of computational methods, such as simple SAR and QSAR, as well as methods for chemical grouping and formalized approaches based on chemical similarity analysis.<sup>9</sup> (Q)SAR methodology consists of a representation of the chemical structure using molecular descriptors and a learning algorithm that relates biological activity of a compound to its chemical structure.<sup>10</sup> While a variety of molecular parameters can be used in the computational methods for (Q)SAR analysis,<sup>10–12</sup> some of these parameters are complex physicochemical or geometrical 3D descriptors whose calculation is associated with difficulties conditioned by molecular flexibility and adequate sampling of conformational space. Conversely, topological indices, or 2D descriptors, obtainable from the structural formula of a compound are very attractive because of their simplicity. A reasonable compromise between ease of interpretation and ease of computation was reported by Carhart et al.,<sup>11</sup> who introduced atom pair descriptors as features of the atomic environments of all pairs of atoms in the 2D representation of a chemical structure. Although this methodology is rather simple, only a few papers have been published where descriptors of this kind were applied in SAR analysis.<sup>12,13</sup>

Among the known serine protease inhibitors, QSAR analysis has been performed for inhibitors of several proteases,<sup>14–16</sup> including the analysis of peptide inhibitors of porcine pancreatic elastase.<sup>17</sup> However, there are currently no reported (Q)SAR models for non-peptide inhibitors of human neutrophil elastase. Here, we utilized computational SAR analysis of a large group of *N*-benzoylpyrazoles to further optimize these molecules as lead neutrophil elastase inhibitors. We present an improved approach to SAR methodology based on atom pair descriptors in combination with classical physicochemical and geometrical descriptors and show that this methodology can detect specific combinations of substructure patterns that confer high or low inhibitory activity against neutrophil elastase. Furthermore, we suggest that the SAR approach developed here may be widely applicable to qualitative SAR recognition problems in other data sets.

## 2. Results and discussion

### 2.1. Descriptors

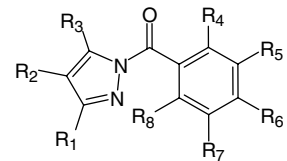
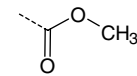
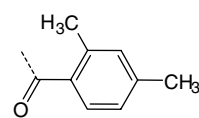
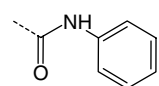
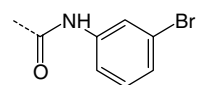
In our investigation, we used atom pairs automatically generated directly from bond connectivity of *N*-benzoylpyrazoles **1–53** (Table 1), as well as physicochemical and structural descriptors obtained from semiempirical calculations and from formulae of the compounds. Atom pairs have been used previously in SAR modeling<sup>11–13,18</sup> and have been defined using different schemes. These differences in nomenclature consisted mainly in the description of atom types. For example, Carhart et al.<sup>11</sup> defined an atom type with its chemical name, number of attached non-hydrogen atoms, and its number of bonding  $\pi$ -electrons. A similar definition of atom types was used by Rusinko

et al.<sup>13</sup> in their analysis of chemical libraries. In comparison, Seierstad and Agrafiotis<sup>18</sup> described atom types in terms of the SMARTS definition, taking into account hydrogen bond donor/acceptor, lipophilicity, and charge characteristics instead of explicit chemical notation. In most of these papers the atom pairs were used as indicator variables with Boolean values (0 or 1).

Various conventions for naming atom types are implemented in molecular modeling software. For example, MM+, AMBER, OPLS and other force fields developed for molecular mechanics computations assign very specific names to atoms of the same chemical nature depending on their environment in a molecule. For SAR analysis, we used the atom naming scheme 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 53 *N*-benzoylpyrazoles (Fig. 1). These frequencies were considered as values of the corresponding atom pair descriptors. Hence, the atom pairs used had non-indicator character, which is another distinctive feature of our approach. This characteristic has some advantages over Boolean values used previously in trend-vector analysis<sup>11</sup> and recursive partitioning methods.<sup>13</sup> For example, cyclohexane and cycloheptane have the same set of atom pairs (C4\_1\_C4, C4\_2\_C4, C4\_3\_C4), which would not be distinguished from each other using Boolean (indicator) values of descriptors. However, the frequencies of occurrence of these atom pairs are different in six- and seven-membered rings, resulting in different values of non-indicator descriptors between two cyclic hydrocarbons, as well as between their derivatives.

Several 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, C4\_1\_CA corresponds to the methylated aromatic ring (shown in red for Compounds **11** and **14**), C3\_4\_C4 corresponds to the presence of a methyl group within a substituent of the pyrazole moiety (shown in red for Compound **2**), and C4\_4\_C4 represents two methyl groups present as R<sub>1</sub> and R<sub>3</sub> substituents of the heterocycle (shown in red for Compounds **5** and **50**). Atom pairs C3\_1\_NO and NO\_1\_O1 both correspond to nitro groups (shown in bold for Compounds **30** and **32**) and can be regarded as having the same chemical origin. However, they are not completely equivalent, as NO\_1\_O1 represents any nitro group, including substituents in aromatic moieties, while C3\_1\_NO designates only nitro groups in the pyrazole ring. Four atom pairs (C3\_1\_C3, C3\_1\_N2, N2\_1\_N2, C3\_2\_C3) had equal occurrences in compounds **1–53**, as they originate from the pyrazole ring and are present in all of the compounds investigated. Thus, such descriptors with zero variance were excluded from further consideration. It should be noted that, although atom naming was taken

**Table 1.** Structure and neutrophil elastase inhibitory activity of *N*-benzoylpyrazoles<sup>a</sup>

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>	R <sub>8</sub>	K <sub>i</sub> (nM)
									
1	H	Cl	H	H	H	F	H	H	6
2	H	H		H	H	Cl	H	H	15
3	CH <sub>3</sub>	H	H	H	OCH <sub>3</sub>	OCH <sub>3</sub>	OCH <sub>3</sub>	H	21
4	H	Cl	H	H	NO <sub>2</sub>	H	H	H	24
5		H	H	H	H	F	H	H	24
6	NO <sub>2</sub>	H	H	H	H	CH <sub>3</sub>	H	H	28
7	H	NO <sub>2</sub>	H	H	H	H	H	H	34
8		H	H	H	H	H	H	H	39
9	H	Br	H	H	H	CH <sub>3</sub>	H	H	45
10	NO <sub>2</sub>	H	H	H	H	H	H	H	46
11	H	NO <sub>2</sub>	H	CH <sub>3</sub>	H	H	H	H	65
12	H	H	H	H	Cl	Cl	H	H	104
13	NO <sub>2</sub>	H	CH <sub>3</sub>	F	H	H	H	H	107
14	H	Cl	H	CH <sub>3</sub>	H	H	H	H	230
15	H	Br	H	H	CH <sub>3</sub>	H	H	H	250
16	H	Br	H	CH <sub>3</sub>	H	H	H	H	300
17	CH <sub>3</sub>	H	H	H	H	NHCOCH <sub>3</sub>	H	H	300
18	H	Cl	H	Cl	H	H	H	H	1000
19	H	Br	H	F	H	H	H	F	1100
20	H	H	H	CH <sub>3</sub>	H	H	H	H	3400
21	H	Br	H	H	F	F	H	Cl	7200
22	H	Cl	H	H	H	<i>tert</i> -Butyl	H	H	9000
23	CH <sub>3</sub>	Cl	CH <sub>3</sub>	F	H	H	H	H	9000
24	CH <sub>3</sub>	Cl	CH <sub>3</sub>	H	H	Cl	H	H	10700
25	CH <sub>3</sub>	Br	CH <sub>3</sub>	H	H	OCH <sub>3</sub>	H	H	24500
26	H	H	H	Br	H	H	H	H	29900
27	CH <sub>3</sub>	CH <sub>3</sub>	CH <sub>3</sub>	H	OCH <sub>3</sub>	OCH <sub>3</sub>	OCH <sub>3</sub>	H	50900
28	H	H	H	H	H	H	H	Cl	NA <sup>b</sup>
29	H	H	H	H	F	H	H	Cl	NA
30	H	NO <sub>2</sub>	H	F	H	H	H	H	NA
31	H	NO <sub>2</sub>	H	H	Br	H	H	H	NA
32	H	NO <sub>2</sub>	H	Cl	H	Cl	H	H	NA
33		H	H	Cl	H	H	H	H	NA
34	CH <sub>3</sub>	H	CH <sub>3</sub>	H	H	Cl	H	H	NA
35	CH <sub>3</sub>	H	CH <sub>3</sub>	H	H	<i>tert</i> -Butyl	H	H	NA
36	CH <sub>3</sub>	H	CH <sub>3</sub>	H	NO <sub>2</sub>	Cl	H	H	NA
37	CH <sub>3</sub>	H	CH <sub>3</sub>	H	Br	CH <sub>3</sub>	H	H	NA
38	CH <sub>3</sub>	H	CH <sub>3</sub>	H	H	NHCOCH <sub>3</sub>	H	H	NA
39	CH <sub>3</sub>	H	CH <sub>3</sub>	H	OCH <sub>3</sub>	OCH <sub>3</sub>	H	H	NA
40	CH <sub>3</sub>	H	CH <sub>3</sub>	OCH <sub>3</sub>	H	H	H	H	NA
41	CH <sub>3</sub>	H	CH <sub>3</sub>	H	Cl	H	H	H	NA
42	CH <sub>3</sub>	H	CH <sub>3</sub>	COOH	H	H	H	H	NA
43	CH <sub>3</sub>	H	CH <sub>3</sub>	H	H	OCH <sub>3</sub>	H	H	NA
44	CH <sub>3</sub>	H	CH <sub>3</sub>	Cl	H	H	H	H	NA
45	CH <sub>3</sub>	H	CH <sub>3</sub>	H	CH <sub>3</sub>	NO <sub>2</sub>	H	H	NA

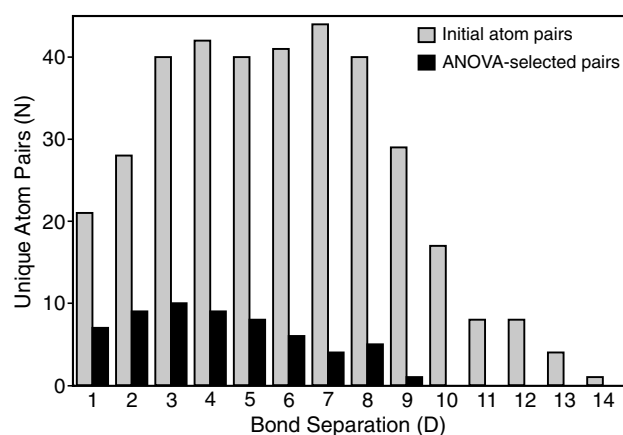
(continued on next page)

Table 1 (continued)

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>	R <sub>8</sub>	K <sub>i</sub> (nM)
46	CH <sub>3</sub>	H	CH <sub>3</sub>	H	H	NHCOCH <sub>2</sub> CH <sub>3</sub>	H	H	NA
47	CH <sub>3</sub>	Cl	CH <sub>3</sub>	Cl	H	Cl	H	H	NA
48	CH <sub>3</sub>	Cl	CH <sub>3</sub>	H	Cl	Cl	H	H	NA
49	CH <sub>3</sub>	Cl	CH <sub>3</sub>	Br	H	H	H	H	NA
50	CH <sub>3</sub>	Cl	CH <sub>3</sub>	H	H	Cl	NO <sub>2</sub>	H	NA
51	CH <sub>3</sub>	Br	CH <sub>3</sub>	H	Br	H	H	H	NA
52	CH <sub>3</sub>	Br	CH <sub>3</sub>	H	H	Cl	H	H	NA
53	CH <sub>3</sub>	H	CH <sub>3</sub>	H	Cl	Cl	H	H	NA

<sup>a</sup> Data taken from Ref. 7.

<sup>b</sup> NA, not active or no inhibition seen at the highest concentration of compound tested (55  $\mu$ M).



**Figure 1.** Numbers of unique atom pairs in 53 *N*-benzoylpyrazoles. The numbers are shown for each of the indicated bond separations initially generated for the 53 *N*-benzoylpyrazoles (light bars). Atom pairs subsequently selected by ANOVA as having significant differences between the three classes of elastase inhibitory activity are shown in dark bars.

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. Thus, initial data for each compound might include just a sketch of the molecular formula saved in HIN format. Nevertheless, we did perform geometry optimization here with MM+ force field and then by the semi-empirical PM3 method for the purpose of physicochemical descriptor calculations by HyperChem.

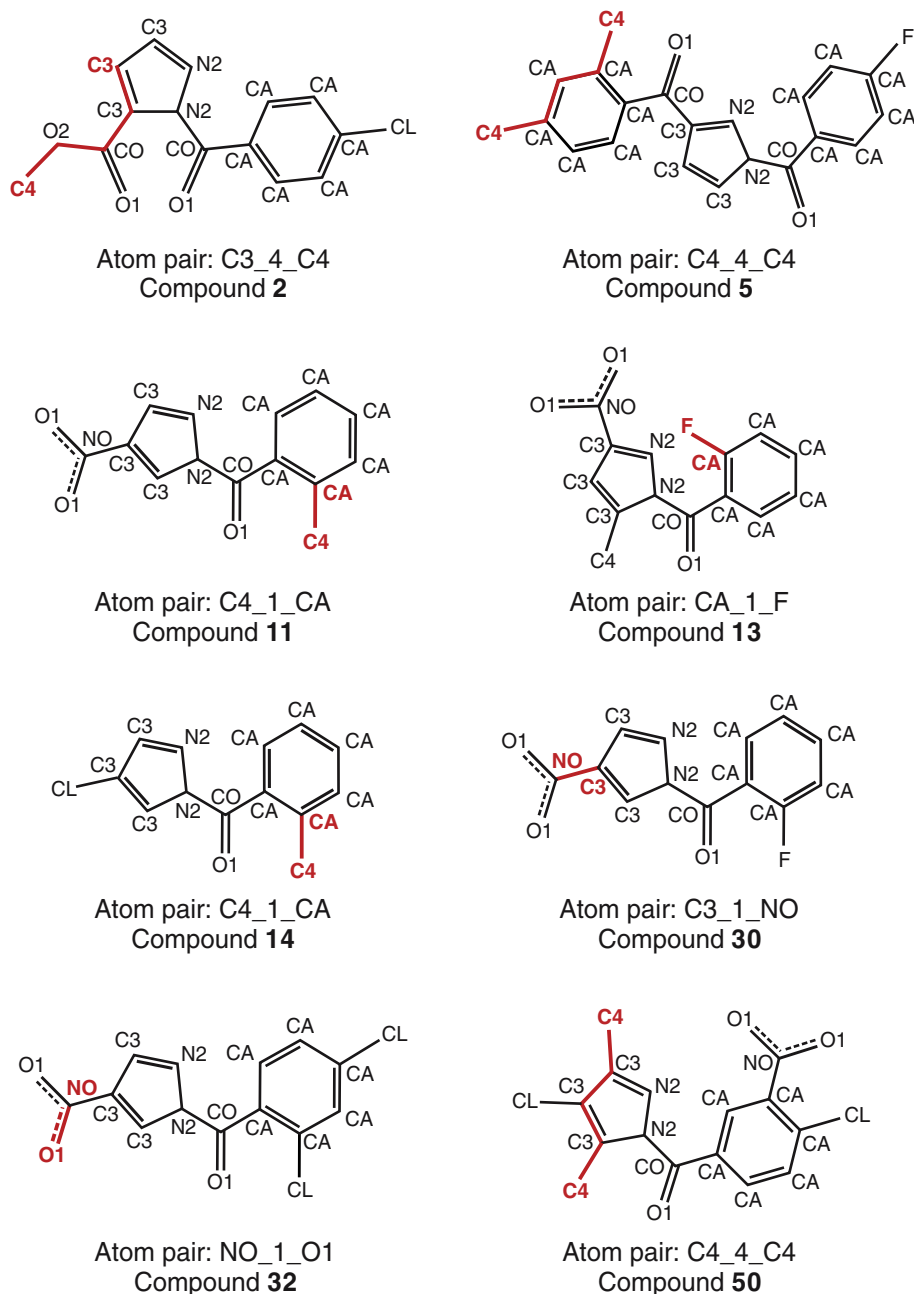
## 2.2. SAR modeling

The total set of descriptors used can be divided into atom pairs, 2D descriptors obtained directly from structural formulae of the compounds, and physicochemical descriptors. The latter group of descriptors required application of additional techniques for determination, such as semi-empirical methods for orbital energy calculation and procedures for evaluation of lipophilicity and molar refractivity (see Section 4). Hence, we performed two separate SAR analyses for comparison, one based on the entire set of descriptors, and the other based on the set of atom pairs and 2D parameters only. If results of comparable quality could be obtained for these two

types of analyses, then the less complex methodology without physicochemical characteristics would be beneficial because this approach allows easy visualization and translation of these variables into a simple ‘chemical’ language, as the variables correspond directly to the presence of certain chemical substructures and functional groups in a molecule.

SAR analysis using atom pairs, quantum-mechanical characteristics calculated by the semi-empirical PM3 method, ‘classical’ physicochemical descriptors (*Refr*, *Refr*(Pz), *Refr*(Ph), ACD/Log *P*, *E*<sub>HOMO</sub>, *E*<sub>LUMO</sub>), and integer variables obtained directly from structural formulae (*n*<sub>1</sub>, *n*<sub>2</sub>, *n*<sub>3</sub>, *n*<sub>o</sub>, *n*<sub>m</sub>, *n*<sub>p</sub>) resulted in an initial data matrix of descriptors that contained 375 columns (variables) (see Section 4 for variable definitions). The second analysis utilizing atom pairs and integer variables only resulted in an initial data matrix of descriptors containing 369 variables. Large numbers of variables requires selection of descriptors to obtain a shortened list of molecular characteristics that are the most valid for effective SAR analysis. For example, the method of recursive partitioning can be applied when the number of variables is quite large, consisting of thousands or even millions of descriptors; however, this method does not account well for possible covariance of variables and uses a simplified descriptor selection with construction of a hierarchical tree.<sup>13,19,20</sup> In comparison, the more elaborate procedure of linear discriminant analysis (LDA) can account for clustering data in multidimensional space of interrelated variables, but cannot be accomplished if there are too many descriptors. In order to exploit the advantages of LDA, we performed a sequential variable selection with application of a specified procedure on each stage.

ANOVA methodology was used in the first stage of selection to identify variables most important for distinguishing compounds in descriptor space.<sup>21</sup> The set of *N*-benzoylpyrazoles was divided into three classes based on elastase inhibitory activity: High, Medium, and not-active (NA) (see Section 4). We performed ANOVA calculation of significance for differences between means in these three classes for each of the 375 initial columns of data matrix, and this procedure resulted in selection of 66 descriptors having significant (*p* < 0.05) difference between within-class and total variability. This group of parameters contained four physicochemical descrip-



**Figure 2.** Examples of atom pairs from the best subsets in the structures of representative *N*-benzoylpyrazoles. Atom pairs are depicted in red and indicated below the structure. Compound numbers correspond to those shown in Table 1.

tors [*Refr*, *ACD/LogP*, *E<sub>LUMO</sub>*, *Refr(Ph)*], three 2D integer variables ( $n_1$ ,  $n_3$ ,  $n_6$ ), and 59 atom pairs. Figure 1 shows relative number of atom pairs with different bond separations selected by ANOVA and initially generated by CHAIN for the 53 compounds. Atom pairs separated by 2–4 bonds seemed to be the most important for detailed SAR analysis because they better discriminated between the three activity classes (Fig. 1). Only one atom pair descriptor with a 9-bond separation (CA\_9\_O1) was selected by ANOVA for further SAR analysis; whereas, atom pairs separated by  $\geq 10$  chemical bonds were not significant, and differences of their occurrence in three classes were negligible at a confidence level of  $p < 0.05$ .

In the second stage of sequential variable selection, the LDA procedure was applied to the 66 descriptors chosen by ANOVA, since reduction of the dimension of descriptor space from 375 to 66 variables allowed analysis with this powerful methodology. At this point, the SAR calculations were branched into two routes. Route 1 LDA utilized all 66 descriptors, including physico-chemical characteristics; whereas, Route 2 analysis included only atom pairs and 2D descriptors obtained directly from molecular graphs of *N*-benzoylpyrazoles. In the latter case, *Refr*, *ACD/LogP*, *E<sub>LUMO</sub>*, and *Refr(Ph)* were discarded, and 62 variables were retained as an input for LDA. The implementation of LDA in STATISTICA 6.0 made it possible to sort out variables



which were non-significant for SAR classification by this method, as the corresponding coefficients of classification functions were automatically zeroed by STATISTICA 6.0. LDA led to a further decrease in the number of important descriptors, and the resulting LDA classification matrices are shown in Table 2. LDA using Route 1 retained 25 variables [*Refr*, ACD/Log *P*, *E*<sub>LUMO</sub>, *Refr*(Ph), *n*<sub>1</sub>, *n*<sub>3</sub>, *n*<sub>o</sub>, BR\_1\_C3, C3\_1\_C4, C3\_1\_CO, C3\_1\_NO, C4\_1\_CA, CA\_1\_F, NO\_1\_O1, C3\_2\_C4, C4\_2\_N2, C3\_3\_C4, C4\_3\_CO, C4\_3\_N2, C3\_4\_C4, C4\_4\_C4, C4\_4\_CO, C4\_4\_O1, C4\_5\_CA, C4\_8\_CL]. In comparison, Route 2 retained the same set of descriptors, with the exception of the four physicochemical variables that were initially discarded (21 variables retained). Atom pairs with one-bond separation (7 of the retained variables) prevailed among all the atom pairs selected by LDA, although several atom pair descriptors with 3-bond and 4-bond separation were also present (3 and 4 of the retained variables, respectively). Table 2 shows that quality of SAR classification for the 53 *N*-benzoylpyrazoles using 25 descriptors (Route 1) was sufficiently high, and the correct activity classification was calculated for 11 of 13 active, all 10 moderately active, and 29 of 30 inactive *N*-benzoylpyrazoles. In total, 50 compounds (94.3%) were classified correctly. Discarding physicochemical characteristics from the descriptor set in Route 2 resulted in a slight decrease in quality of the SAR analysis, and correct activity classification was obtained for 47 of 53 compounds (88.7% accuracy) (Table 2). Estimation of predictive power of the LDA models made by the leave-one-out (LOO) approach is shown in Supplementary Table S1 for Routes 1 and 2, where calculated classes for each compound are also presented. For both routes, *a priori* LOO classification gave 34 coincidences (64.2%) between experimental and predicted classes. While this percentage is low, the number of descriptors used (25 and 21 for Routes 1 and 2, respectively) is still rather high for LOO prediction of biological activity. Thus, a further reduction in the number of descriptors was needed.

The third stage of sequential variable selection utilized the ‘best subset search’ option of LDA, as implemented in STATISTICA 6.0. This option allows one to check possible combinations of descriptors and obtain a smaller combination that is optimal in terms of analysis misclassification or cross-validation misclassification. Because the number of variables was still high, the LDA with ‘best subset search’ was preceded by first removing correlated descriptors from the variable sets

retained after classical LDA on Routes 1 and 2 (see Section 4). Hence, 15 descriptors were retained for LDA with ‘best subset search’ using Route 1 [*Refr*, ACD/Log *P*, *E*<sub>LUMO</sub>, *Refr*(Ph), *n*<sub>1</sub>, *n*<sub>o</sub>, BR\_1\_C3, C3\_1\_CO, C3\_1\_NO, C4\_1\_CA, CA\_1\_F, NO\_1\_O1, C3\_4\_C4, C4\_4\_C4, C4\_8\_CL], and 11 of these variables were retained for analysis using Route 2 (as above, the physicochemical descriptors were excluded for Route 2). Using LDA with the ‘best subset search’ option, we found that the optimal subsets, in terms of analysis misclassification, consisted of 7 and 6 variables for Routes 1 and 2, respectively. Linear classification functions for elastase inhibitory activity classes of High, Medium, and NA are provided below.tpb

#### Route 1:

$$F_{\text{High}} = -51.068 + 1.596\text{Refr} - 17.351n_1 + 3.688n_o + 14.793\text{C3}_1\text{NO} - 4.965\text{C4}_1\text{CA} - 13.687\text{C3}_4\text{C4} + 9.722\text{C4}_4\text{C4} \quad (1)$$

$$F_{\text{Medium}} = -53.379 + 1.628\text{Refr} - 20.507n_1 + 7.365n_o + 9.699\text{C3}_1\text{NO} - 3.341\text{C4}_1\text{CA} - 22.905\text{C3}_4\text{C4} + 12.687\text{C4}_4\text{C4} \quad (2)$$

$$F_{\text{NA}} = -64.130 + 1.798\text{Refr} - 23.782n_1 + 6.372n_o + 15.213\text{C3}_1\text{NO} - 7.462\text{C4}_1\text{CA} - 24.375\text{C3}_4\text{C4} + 21.205\text{C4}_4\text{C4} \quad (3)$$

#### Route 2:

$$F_{\text{High}} = -3.041 + 0.700n_o + 1.323\text{C4}_1\text{CA} + 0.699\text{CA}_1\text{F} + 1.795\text{NO}_1\text{O1} + 4.944\text{C3}_4\text{C4} + 1.004\text{C4}_4\text{C4} \quad (4)$$

$$F_{\text{Medium}} = -4.473 + 3.657n_o + 3.461\text{C4}_1\text{CA} + 1.529\text{CA}_1\text{F} - 0.198\text{NO}_1\text{O1} - 3.504\text{C3}_4\text{C4} + 1.899\text{C4}_4\text{C4} \quad (5)$$

$$F_{\text{NA}} = -4.668 + 4.006n_o - 0.367\text{C4}_1\text{CA} - 1.351\text{CA}_1\text{F} + 1.412\text{NO}_1\text{O1} - 1.137\text{C3}_4\text{C4} + 8.206\text{C4}_4\text{C4} \quad (6)$$

According to SAR rules expressed by these equations, a compound will be assigned to a certain activity class if the value of its corresponding classification function is greater than the values of the functions for the two remaining classes. The results presented in Tables 3

**Table 2.** Classification matrices for linear discriminant analysis-derived SAR with 25 variables (Route 1) and 21 variables (Route 2)

Experimentally determined classification	Calculated classification							
	Route 1				Route 2			
	High	Medium	NA	Accuracy (%)	High	Medium	NA	Accuracy (%)
High	<b>11</b> <sup>a</sup>	1	1	84.6	<b>11</b>	2	0	84.6
Medium	0	<b>10</b>	0	100.0	0	<b>8</b>	2	80.0
NA	0	1	<b>29</b>	96.7	1	1	<b>28</b>	93.3
Total	11	12	30	94.3	12	11	30	88.7

<sup>a</sup> The number of correctly classified compounds is indicated in bold.

**Table 3.** Classification matrices for linear discriminant analysis-derived SAR with best subsets of 7 variables (Route 1) and 6 variables (Route 2)

Experimentally determined classification	Calculated classification							
	Route 1				Route 2			
	High	Medium	NA	Accuracy (%)	High	Medium	NA	Accuracy (%)
High	<b>12</b> <sup>a</sup>	1	0	92.3	<b>12</b>	1	0	92.3
Medium	1	<b>8</b>	1	80.0	1	<b>7</b>	2	70.0
NA	0	3	<b>27</b>	90.0	2	1	<b>27</b>	90.0
Total	13	12	28	88.7	15	9	29	86.8

<sup>a</sup> The number of correctly classified compounds is indicated in bold.

and 4 indicate that SAR classifications with similar qualities were achieved for both Routes 1 and 2 (88.7% and 86.8% of the compounds were correctly calculated with the experimentally determined classes of elastase inhibitory activity, respectively). Thus, inclusion of physicochemical descriptors provided little improvement in the SAR results. Although the quality of LDA with ‘best subsets search’ was generally lower than that obtained by standard LDA (Table 2), use of the ‘best subsets search’ with fewer descriptors led to a significantly higher percentage of correct LOO predictions (Table 4). Indeed, accuracy of *a priori* LOO predictions was 75.5% and 71.7% for Routes 1 and 2, respectively, which corresponded to 40 and 38 of the 53 compounds correctly assigned to their experimental classes using training sets each consisting of 52 *N*-benzoylpyrazole derivatives. The previous stage of sequential variable selection resulted in only 64% correct LOO classifications (Supplementary Table S1). Thus, step-by-step reduction in the number of descriptors led to variable subsets that were the most critical for SAR analysis. Importantly, the resulting classification functions, expressed by Eqs. 1–3 for Route 1 and Eqs. 4–6 for Route 2 appear to fairly accurately predict the class of elastase inhibitory activity of a given *N*-benzoylpyrazole.

Coefficients of linear classification functions (see Eqs. 1–6) reflect cooperative effects of partially correlated descriptors on the values of  $F_{\text{High}}$ ,  $F_{\text{Medium}}$ ,  $F_{\text{NA}}$ , and direct interpretation of their values is complex; however, unambiguous interpretation of these values was sometimes possible. For example, values for variables  $n_o$  and C4\_4\_C4 tended to following the sequence NA > Medium > High (Fig. 3). Consequently, higher values of these descriptors lead to larger increments for  $F_{\text{NA}}$  than for  $F_{\text{High}}$  and are unfavorable for elastase inhibitory activity.

Note that other robust (Q)SAR methods exist, which provide good fitting and prediction results without the use of variable selection. Among the best of these approaches is the random forest method.<sup>22</sup> However, this method is not able to produce an explicit model from randomly generated trees. Rather, the random forest model can be regarded as a ‘black box’.<sup>22</sup> Since the goal of our approach was to construct interpretable and defined SAR models to predict elastase inhibitory activity of *N*-benzoylpyrazoles, pre-selection of variables based on activity was necessary. This resulted in the derivation

of very simple and intuitively understandable SAR rules, as described below.

### 2.3. Classification tree analysis and derivation of simplified SAR rules

Although Eqs. 1–6 are capable of fitting and predicting elastase inhibitory activity classes of *N*-benzoylpyrazoles with good accuracy, simpler and intuitively understandable SAR criteria are desirable. We tried to reasonably simplify classification rules with the use of binary classification tree analysis, as implemented in STATISTICA 6.0. Based on 7 descriptors from the best subset selected on Route 1, we have built the optimal classification tree shown in Figure 4. The tree has three splits according to the three conditions indicated for the splits. If a condition is satisfied, then the compounds are sent to the left branch of the tree, otherwise they are sent to the right branch, eventually reaching one of the terminal nodes corresponding to a certain activity class. For example, the condition C4\_4\_C4 < 0.38 indicates that compounds not containing atom pairs of this type (i.e., when the integer value of the C4\_4\_C4 descriptor is equal to zero), will be sent to the left branch (28 cases), while compounds containing one or more C4\_4\_C4 atom pairs will be sent to the right branch and immediately enter the NA terminal node. Thus, according to the above-mentioned observation from LDA analysis (Fig. 3B), the presence of C4\_4\_C4 in the chemical structure of an *N*-benzoylpyrazole is unfavorable for inhibitory activity, and such compounds should be designated inactive. Note that in almost all the compounds investigated, this atom pair is associated with two methyl groups R<sub>1</sub> and R<sub>3</sub> in the pyrazole moiety (see an example of this atom pair within the structure of Compound 50 in Fig. 2). The only exception to this rule was Compound 5, where the C4\_4\_C4 descriptor originates from two methyl substituents in the phenyl ring (Fig. 2).

The logic tree for Route 1 is defined by two additional descriptors from the best subset, namely by the conditions C3\_1\_NO < 0.64 and  $n_o$  < 0.42 (Fig. 4). Integer values of descriptors participating in these conditions indicates that the presence of a C3\_1\_NO atom pair (i.e., nitro group substituent of the pyrazole ring; see an example of this atom pair in compound 30 of Fig. 2) or *ortho*-substituents in the benzoyl moiety sends compounds to the right branches on the two

**Table 4.** Results of SAR classification and leave-one-out (LOO) prediction obtained using Routes 1 and 2 for each compound based on LDA with the ‘best subset search’ option

Compound	Experimental classification	Route 1		Route 2	
		Calculated class	LOO-predicted class	Calculated class	LOO-predicted class
1	High	High	<i>Medium<sup>a</sup></i>	High	<i>Medium</i>
2	High	High	High	High	High
3	High	High	High	High	High
4	High	High	<i>Medium</i>	High	High
5	High	High	High	High	<i>Medium</i>
6	High	High	High	High	High
7	High	High	<i>NA</i>	High	High
8	High	High	High	High	High
9	High	<i>Medium</i>	<i>Medium</i>	<i>Medium</i>	<i>Medium</i>
10	High	High	High	High	High
11	High	High	<i>NA</i>	High	<i>Medium</i>
12	High	High	<i>Medium</i>	High	High
13	High	High	High	High	High
14	Medium	Medium	Medium	Medium	Medium
15	Medium	Medium	Medium	Medium	<i>High</i>
16	Medium	Medium	Medium	Medium	Medium
17	Medium	<i>High</i>	<i>High</i>	<i>High</i>	<i>High</i>
18	Medium	Medium	Medium	<i>NA</i>	<i>NA</i>
19	Medium	Medium	Medium	Medium	Medium
20	Medium	Medium	Medium	Medium	Medium
21	Medium	Medium	Medium	Medium	Medium
22	Medium	Medium	Medium	Medium	<i>High</i>
23	Medium	<i>NA</i>	<i>NA</i>	<i>NA</i>	<i>NA</i>
24	NA	NA	NA	NA	NA
25	NA	NA	NA	NA	NA
26	NA	<i>Medium</i>	<i>Medium</i>	NA	<i>Medium</i>
27	NA	NA	NA	NA	NA
28	NA	<i>Medium</i>	<i>Medium</i>	NA	<i>Medium</i>
29	NA	<i>Medium</i>	<i>Medium</i>	<i>Medium</i>	<i>Medium</i>
30	NA	NA	NA	<i>High</i>	<i>High</i>
31	NA	NA	<i>High</i>	<i>High</i>	<i>High</i>
32	NA	NA	NA	NA	NA
33	NA	NA	<i>Medium</i>	NA	<i>Medium</i>
34	NA	NA	NA	NA	NA
35	NA	NA	NA	NA	NA
36	NA	NA	NA	NA	NA
37	NA	NA	NA	NA	NA
38	NA	NA	NA	NA	NA
39	NA	NA	NA	NA	NA
40	NA	NA	NA	NA	NA
41	NA	NA	NA	NA	NA
42	NA	NA	NA	NA	NA
43	NA	NA	NA	NA	NA
44	NA	NA	NA	NA	NA
45	NA	NA	NA	NA	NA
46	NA	NA	NA	NA	NA
47	NA	NA	NA	NA	NA
48	NA	NA	NA	NA	NA
49	NA	NA	NA	NA	NA
50	NA	NA	NA	NA	NA
51	NA	NA	NA	NA	NA
52	NA	NA	NA	NA	NA
53	NA	NA	NA	NA	NA

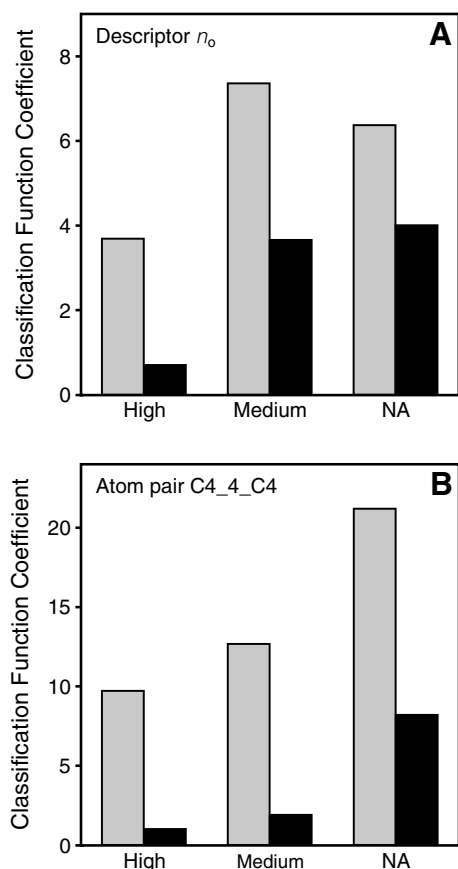
<sup>a</sup> Incorrect classifications are indicated in italics.

corresponding logical splits of the tree, because any non-zero positive values of C3\_1\_NO and  $n_o$  violate these conditions, respectively. [Supplementary Table S2](#) shows SAR classification for each of the 53 *N*-benzoylpyrazoles analyzed by the classification tree shown in [Figure 4](#). Taking into account a structural sense of

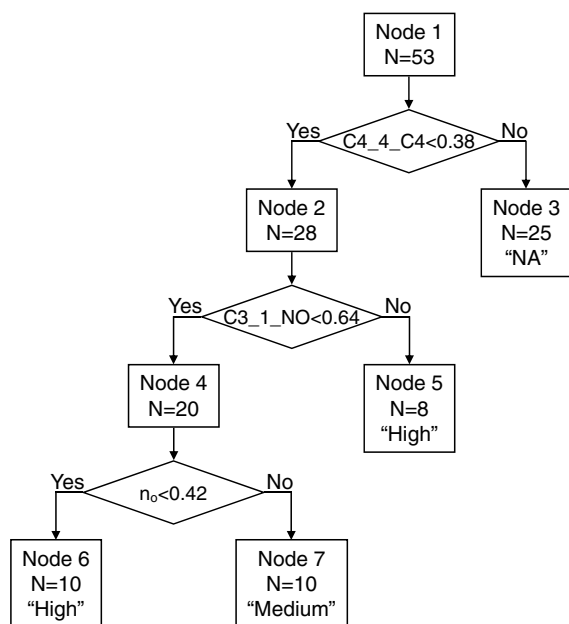
the atom pairs ([Fig. 2](#)), this classification tree can also be presented in the form of a simple ‘chemical’ algorithm ([Scheme 1](#)).

Note that these rules are not completely equivalent with the classification tree ([Fig. 4](#)), since the C4\_4\_C4





**Figure 3.** Classification function coefficients for selected descriptors. Coefficients were obtained using the best subsets of 7 variables (Route 1, light bars) and 6 variables (Route 2, dark bars) for descriptor  $n_o$  (A) and atom pair C4\_4\_C4 (B).



**Figure 4.** Classification tree for predicting elastase inhibitory properties of *N*-benzoylpyrazoles using Route 1. The number of *N*-benzoylpyrazoles that entered each node is indicated. Terminal nodes correspond to the three activity classes: High, Medium, or non-active (NA).

atom pair can also be associated with methyl groups located in the phenyl ring, rather than the pyrazole moiety (see Fig. 2, compound 5). Using the tree, compound 5 is classified incorrectly; whereas, this *N*-benzoylpyrazole derivative is classified correctly by the ‘chemical’ algorithm, which considers methyl substituents only in the pyrazole moiety. In total, the correct elastase inhibitory activity class was determined by Scheme 1 for 42 of the 53 *N*-benzoylpyrazoles (79.2% accuracy). While LDA with ‘best subset search’ resulted in a higher percentage (88.7%) of correctly calculated classifications (Table 3), the simplified SAR rules are attractive because of their clarity and the possibility of translating them into standard ‘chemical’ language.

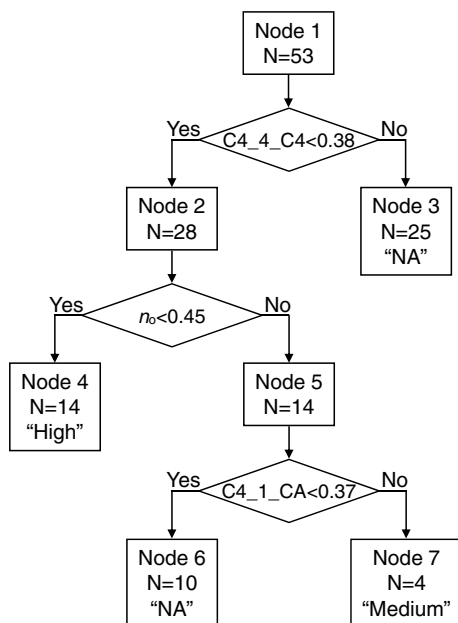
The classification tree approach was also applied to the best subset of 6 variables selected by LDA on Route 2. An optimal tree obtained in this case considered 3 variables (Fig. 5). Again, descriptors C4\_4\_C4 and  $n_o$  were important in the derived SAR rules. The third condition was based on the C4\_1\_CA atom pair, which is associated with the presence of alkyl groups attached to the aromatic ring (see Compounds 11 and 14 in Fig. 2). In contrast, the C3\_1\_NO descriptor involved in the classification tree for Route 1 was not present in the best LDA subset identified with Route 2 (Eqs. 4–6). Supplementary Table S2 shows SAR classification for each of the 53 *N*-benzoylpyrazoles according to the tree shown in Figure 5. This tree can also be presented in the form of a simple ‘chemical’ algorithm (Scheme 2). Hence, according to the SAR rules derived with Route 2, a compound is classified as inactive if it has non-methyl *ortho*-substituents (i.e., Cl, Br, F, etc.) or two methyl substituents in the pyrazole moiety. Likewise, highly active inhibitors cannot contain two methyl groups in the pyrazole heterocycle or *ortho*-substituents in benzoyl radical. The Scheme 2 algorithm correctly classified 43 of 53 *N*-benzoylpyrazoles (81.1%), which is comparable to the 79.2% accuracy obtained on Route 1 using SAR rules from Scheme 1. In spite of the satisfactory overall percentage of correct classifications, both algorithms had lower accuracy in predicting the activity class of the 10 moderately active compounds (4 and 3 cases on Routes 1 and 2, respectively). Clearly, this is due to the relative simplicity of binary classification trees, as they represent only rough approximations to the more sophisticated LDA-derived classifications expressed by Eqs. 1–6.

The simplified SAR rules did not differ significantly between Routes 1 and 2 in quality of their results. Consequently, inclusion of physicochemical characteristics in a descriptor set does not seem to provide a significant advantage. The easily calculated 2D descriptors (atom pairs and simple structural variables) used on Route 2 were sufficient to obtain good accuracy for SAR recognition, both by LDA and by binary classification tree analysis. Importantly, these simplified SAR rules were generally consistent with our molecular modeling data,<sup>7</sup> where we found that R<sub>1</sub> and R<sub>3</sub> groups in a pyrazole, as well as *ortho*-substituents in the benzoyl radical

```

If pyrazole moiety contains methyl groups R1 and R3 then the compound is inactive
else
  ( if pyrazole moiety contains a nitro group
    or benzoyl moiety does not contain ortho-substituents
    then the compound is highly active
    else the compound is moderately active )
  
```

Scheme 1.



**Figure 5.** Classification tree for predicting elastase inhibitory properties of *N*-benzoylpyrazoles using Route 2. The number of *N*-benzoylpyrazoles that entered each node is indicated. Terminal nodes correspond to the three activity classes: High, Medium, or non-active (NA).

prevented proper positioning of the *N*-benzoylpyrazole required for interaction between the hydroxyl group of elastase Ser195 making them unfavorable for nucleophilic attack by Ser195 to the carbonyl group of an inhibitor in the oxyanion hole. Likewise, the presence of a nitro group enhanced positive charge on the *N*-benzoylpyrazole carbonyl carbon atom, making it more susceptible to attack by the nucleophile.<sup>7</sup>

### 3. Conclusions

Previously, we utilized high-throughput screening to select unique small-molecule inhibitors of neutrophil elastase, and identified a novel class of elastase inhibitors with an *N*-benzoylpyrazole scaffold.<sup>7</sup> Here, we performed

SAR analysis of these compounds to further define the features of these molecules important for activity and to develop a simple, but accurate SAR model for predicting biological activity in future compound screening. The analysis of structure–activity relationships is an important approach for defining the critical combination of various structural and physicochemical descriptors responsible for the biological activity of a given molecule (e.g.,<sup>23,24</sup>). In the present study, we utilized 2D atom pair descriptors together with physicochemical molecular descriptors (Route 1) or 2D parameters alone (Route 2) for SAR analysis of a series *N*-benzoylpyrazoles with various levels of experimentally determined elastase inhibitory activity.

A sequence of ANOVA, linear discriminant, and binary classification tree analyses based on the molecular descriptors led to the derivation of simple SAR rules, in spite of the large number of starting variables. The SAR rules obtained by binary classification tree analysis on Routes 1 and 2 were quite consistent with our experimental activity data and molecular docking studies,<sup>7</sup> indicating the approach can accurately predict active elastase inhibitors. We also found that physicochemical characteristics (i.e., energies of frontier molecular orbitals, molar refractions, lipophilicities) were not necessary for achieving good SAR rules, as comparable quality of SAR classification was obtained with 2D descriptors only. Thus, the use of atom pair descriptors is a valuable tool for identifying different SAR rules in high-throughput screening data sets and could provide a relatively simple classification useful for *de novo* design of elastase inhibitors with an *N*-benzoylpyrazole scaffold.

Although we applied atom pair descriptors to SAR in a set of related compounds, this approach is also applicable to chemically diverse data sets.<sup>13</sup> We believe that our modification of the method using more specific atom typing and non-biased values of descriptors, in conjunction with sequential variable selection, will also be useful for SAR analysis in a heterogeneous series of compounds, and this issue will be addressed in future studies.

```

If pyrazole moiety contains methyl groups R1 and R3 then the compound is inactive
else
  { if benzoyl moiety does not contain ortho-substituents, then the compound is highly active
    else
      ( if ortho-substituent is a methyl group, then the compound is moderately active
        else the compound is inactive ) }
  
```

Scheme 2.

## 4. Materials and methods

### 4.1. Molecular set

The data set used in this study is a series of 53 *N*-benzoylpyrazoles with different levels of inhibitory activity for human neutrophil elastase. These compounds were selected by high-throughput screening of a 10,000-compound chemolibrary.<sup>7</sup> For SAR analysis, the set of the *N*-benzoylpyrazoles (Table 1) was divided into three activity classes according to their experimentally determined elastase inhibitory activity. Inhibitors possessing  $K_i \leq 200$  nM were regarded as highly active and were placed in the activity class labeled ‘High’ (13 compounds). *N*-Benzoylpyrazoles with moderate activity ( $200 < K_i \leq 10,000$  nM) were placed in the activity class labeled ‘Medium’ (10 compounds). Derivatives with  $K_i > 10,000$  nM considered non-active and placed in the activity class labeled ‘NA’ (30 compounds).

### 4.2. Structure encoding by atom pairs and other 2D descriptors

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 represents the topological distance or number of bonds in the shortest path between these atoms in a structural formula. In our investigation, T1 and T2 were defined with symbolic codes used in HyperChem, Version 7 (Hypercube, Inc., Waterloo, ON, Canada) for atom type representation within MM+ force field. For example, CA, CO, and C3 codes were used for  $sp^2$ -hybridized aromatic, carbonyl, and pyrazole 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. The notation of atom types can be changed, if necessary, based on the force field used. For example, the codes listed above for aromatic, carbonyl, and pyrazole carbons would be altered to CA, C, and CM, respectively, if AMBER instead of MM+ force field was used for HyperChem output. As atom pairs T1\_D\_T2 and T2\_D\_T1 are equivalent, we chose a unified definition with lexicographic order of type substrings (i.e., with  $T1 \leq T2$ ).

All 367 unique atom pairs possible for non-hydrogen atoms in the 53 *N*-benzoylpyrazoles were generated. This  $53 \times 367$  data matrix was automatically built by our CHAIN program, based on HIN files created in HyperChem. By convention, 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. The data matrix obtained in this way for the 53 compounds contained columns with no variance for descriptors C3\_1\_C3, C3\_1\_N2, N2\_1\_N2, C3\_2\_C3, because these atom pairs are present in all the compounds investigated at the same frequency. Thus, the corresponding columns were deleted from the matrix, resulting in a  $53 \times 363$  matrix of atom pair descriptors.

In addition to atom pairs, we selected the following set of 6 additional structural 2D descriptors: number of substituents in *ortho*-( $n_o$ ) and *meta*-( $n_m$ ) positions of the benzene ring; and numbers of substituents  $R^1$ ,  $R^2$ ,  $R^3$ ,  $R^6$  (Table 1) denoted as  $n_1$ ,  $n_2$ ,  $n_3$ ,  $n_p$ , respectively (integer variables). These descriptors were obtained directly from structural formulae of Compounds 1–53.

### 4.3. Physicochemical descriptors

The following 6 physicochemical descriptors were used: total molar refraction (*Refr*), lipophilicity (octanol–water partition coefficient; ACD/Log *P*), energies of the highest occupied and lowest unoccupied molecular orbitals ( $E_{HOMO}$  and  $E_{LUMO}$ , respectively), and sum of refractions for substituents in the pyrazole ( $R^1$ ,  $R^2$ ,  $R^3$ ) and benzene ( $R^4$ – $R^8$ ) rings [*Refr*(Pz) and *Refr*(Ph), respectively]. Energies  $E_{HOMO}$  and  $E_{LUMO}$  were determined by the semi-empirical PM3 method after geometry optimization in HyperChem. The values of *Refr*, *Refr*(Pz), and *Refr*(Ph) were calculated with the QSAR built-in module of HyperChem. Lipophilicities ACD/Log *P* were obtained taken from the site [www.emolecules.com](http://www.emolecules.com). The resulting data matrix of physicochemical and structural descriptors and atom pairs contained 375 columns (variables).

### 4.4. Data processing and derivation of SAR rules

Derivation of SAR classification was accompanied by sequential variable selection and reduction of dimensionality. In order to distinguish between variables significant and non-significant for SAR, we applied one-way analysis of variance (ANOVA)<sup>21</sup> using the STATISTICA 6.0 package (StatSoft, Inc., Tulsa, OK). The variables selected by ANOVA served as basic descriptors for refined classification by LDA, using the corresponding module of STATISTICA 6.0. Redundant or non-significant coefficients of the linear classification functions were automatically zeroed by STATISTICA 6.0.

Although LDA methodology allows analysis with linearly dependent variables, we undertook another step of variable selection and excluded dependent correlated variables, as they originated from the same molecular feature. We found that each of 11 descriptors retained by LDA ( $n_3$ , C3\_1\_C4, C3\_2\_C4, C4\_2\_N2, C3\_3\_C4, C4\_3\_N2, C4\_4\_C4, C4\_4\_CO, C4\_5\_CA, C4\_3\_CO, C4\_4\_O1) had correlation coefficients  $>0.8$  with at least one descriptor within this same group. This correlation can be explained easily because all 11 variables are related to the presence of methyl substituents in the compounds. For example, the tetrahedral  $sp^3$ -carbon (atom type C4) is involved in all atom pairs from this group. Similarly, variable  $n_3$  was also present in this group of correlated descriptors because  $R_3$  is a methyl radical in almost all *N*-benzoylpyrazoles containing a substituent in this position of the pyrazole ring. Thus, it was reasonable to choose one of the mutually correlated descriptors as an independent variable representing the entire group of 11 variables. For this purpose, we selected atom pair C4\_4\_C4 as an independent descriptor, because it had the largest sum of correlation coefficients with the remaining 10 variables from this group. Hence, the

dimension of descriptor space was further reduced by 10 variables. We repeated LDA analysis focusing on the remaining descriptors and applying the ‘best subset search’ option available in STATISTICA 6.0.

For the purpose of ultimate visuality we also built simplified SAR rules with the use of binary classification tree methodology.<sup>25</sup> Starting from variables of the best subsets selected as described above, classification trees were built with STATISTICA 6.0 using discriminant-based univariate splits with estimated prior probabilities and equal misclassification costs for classes.<sup>25,26</sup>

### Acknowledgments

This work was supported in part by Department of Defense grant W9113M-04-1-0001, National Institutes of Health grant RR020185, 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 report consist of 1) the results of SAR classification and leave-one-out (LOO) prediction obtained for each *N*-benzoylpyrazole with the initial (classic) LDA analysis and 2) an illustration of simplified SAR rules based on classification tree analysis for each compound. Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.bmc.2008.01.014](https://doi.org/10.1016/j.bmc.2008.01.014).

### References and notes

- Owen, C. A.; Campbell, E. J. *J. Leukoc. Biol.* **1999**, *65*, 137.
- Dollery, C. M.; Owen, C. A.; Sukhova, G. K.; Krettek, A.; Shapiro, S. D.; Libby, P. *Circulation* **2003**, *107*, 2829.
- Carden, D.; Xiao, F.; Moak, C.; Willis, B. H.; Robinson-Jackson, S.; Alexander, S. *Am. J. Physiol.* **1998**, *275*, H385–H392.
- Barnes, P. J.; Stockley, R. A. *Eur. Respir. J.* **2005**, *25*, 1084.
- Chughtai, B.; O’Riordan, T. G. *J. Aerosol Med.* **2004**, *17*, 289.
- Tremblay, G. M.; Janelle, M. F.; Bourbonnais, Y. *Curr. Opin. Invest. Drugs* **2003**, *4*, 556.
- Schepetkin, I. A.; Khlebnikov, A. I.; Quinn, M. T. *J. Med. Chem.* **2007**, *50*, 4928.
- Andricopulo, A. D.; Montanari, C. A. *Mini. Rev. Med. Chem.* **2005**, *5*, 585.
- Worth, A. P.; Bassan, A.; De, B. J.; Gallegos, S. A.; Netzeva, T.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Eisenreich, S. *SAR QSAR Environ. Res.* **2007**, *18*, 111.
- Buttingsrud, B.; Ryeng, E.; King, R. D.; Alsberg, B. K. *J. Comput. Aided Mol. Des.* **2006**, *20*, 361.
- Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64.
- Gute, B. D.; Basak, S. C. *SAR QSAR Environ. Res.* **2006**, *17*, 37.
- Rusinko, A. 3.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017.
- Greenidge, P. A.; Merette, S. A.; Beck, R.; Dodson, G.; Goodwin, C. A.; Scully, M. F.; Spencer, J.; Weiser, J.; Deadman, J. J. *J. Med. Chem.* **2003**, *46*, 1293.
- Frece, V.; Kabelac, M.; De, N. P.; Pricl, S.; Miertus, S. *J. Mol. Graph. Model* **2004**, *22*, 209.
- Li, X.; Zhang, W.; Qiao, X.; Xu, X. *Bioorg. Med. Chem.* **2007**, *15*, 220.
- Nomizu, M.; Iwaki, T.; Yamashita, T.; Inagaki, Y.; Asano, K.; Akamatsu, M.; Fujita, T. *Int. J. Pept. Protein Res.* **1993**, *42*, 216.
- Seierstad, M.; Agrafiotis, D. K. *Chem. Biol. Drug Des.* **2006**, *67*, 284.
- Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 393.
- Rusinko, A. 3.; Young, S. S.; Drewry, D. H.; Gerritz, S. W. *Comb. Chem. High Throughput Screen* **2002**, *5*, 125.
- Lindman, H. R. *Analysis of Variance in Complex Experimental Designs*; W.H. Freeman & Co., San Francisco, **1974**.
- Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947.
- Tong, W.; Welsh, W. J.; Shi, L.; Fang, H.; Perkins, R. *Environ. Toxicol. Chem.* **2003**, *22*, 1680.
- Raevsky, O. A. *Mini. Rev. Med. Chem.* **2004**, *4*, 1041.
- Breiman, L.; Friedman, J. H.; Olshen, R. A. *Stone C.J.* **1984**.
- Loh, W. Y.; Shih, Y. S. *Statistica Sinica* **1997**, *7*, 815.