Ames mutagenicity, structural alerts of carcinogenicity, Hansch QSAR parameters (ClogP, CMR, MgVol), tumor site concordance/multiplicity, and tumorigenicity rank in NTP 2-year rodent studies

Carr J Smith¹, Thomas A Perfetti², Gene M Ko³ and Rajni Garg⁴

Abstract
Since its inception in 1976, the National Toxicology Program (NTP) has conducted 594 2-year studies on rats and mice by a number of routes of administration including inhalation, feed, drinking water, dermal, and intraperitoneal injection. Of these studies, the results on 470 chemicals were of adequate technical quality to be incorporated into final technical reports. In this study, the 470 chemicals were categorized from 1 to 48 by the level of “clear” neoplastic evidence in male and female rats, and in male and female mice, and given an ordinal rank from 1 to 135 following additional considerations regarding tumor site concordance and tumor multiplicity. The resultant tumorigenicity category score and ordinal rank score were examined for associations with results in the Ames Salmonella mutagenicity assay; presence or absence of structural alerts of carcinogenicity; and three Hansch Quantitative Structure-Activity Relationship (QSAR) parameters, namely, calculated base 10 logarithm of the octanol–water partition coefficient (ClogP), calculated molar refractivity (CMR), and McGowan molecular volume (MgVol). Small molecular volumes were found to be associated with higher levels of tumorigenicity. Whereas lower rather than higher levels of lipophilicity were found to be associated with higher levels of tumorigenicity. Positive Ames test results were positively correlated with overall tumorigenicity and with possession of structural alerts. Since larger organic molecules have more chemical reaction centers, it was not surprising that higher ClogP values were positively correlated with the number of structural alerts. The results from this study demonstrate the ability to devise rational rules for relative tumorigenicity that correlate, in biologically plausible ways, with known parameters of toxicity.

Keywords
NTP, structural elements of carcinogenicity, Hansch molecular parameters, tumorigenicity rank, Ames mutagenicity

Date received: 18 December 2017; accepted: 24 January 2018

Introduction
The National Toxicology Program (NTP) is a branch of the US Department of Health and Human Services. A major current emphasis of NTP is “The Toxicology in the 21st Century: The Role of the National Toxicology Program.”¹ NTP describes this program as follows:

The Role of the National Toxicology Program is to support the evolution of toxicology from a predominantly observational...
NTP’s intent is to expand the scientific basis for making public health decisions on the potential toxicity of environmental agents. Over the history of the NTP testing program, 594 different 2-year animal bioassays have been conducted via different routes of exposure including inhalation, feed, drinking water, intraperitoneal injection, and dermal. Of these studies, the results for 470 chemicals were of adequate technical quality to result in the production of a final technical report.

In the current study, each of the 470 chemicals were categorized from 1 to 48 by the level of “clear” neoplastic evidence in male and female rats, and in male and female mice, and given an ordinal rank from 1 to 135 following additional considerations regarding tumor site concordance and tumor multiplicity. The resultant tumorigenicity category score and ordinal rank score were examined for associations with results in the Ames Salmonella mutagenicity assay; presence or absence of structural alerts of carcinogenicity; and three Hansch Quantitative Structure-Activity Relationship (QSAR) parameters, namely, calculated base 10 logarithm of the octanol–water partition coefficient (ClogP), calculated molar refractivity (CMR), and molecular volume (MgVol).

In the present study, we calculated three important molecular parameters for each of the 470 chemical compounds in the NTP database. These molecular parameters (ClogP,2 MgVol,3,4 and CMR5) represent hydrophobic, electronic, and steric effects of a chemical on its biological activity and are extremely useful in developing QSAR models to investigate the quantitative relationship between the biological activity of chemicals and their hydrophobic, electronic, steric, and other physical and chemical characteristics.6

NTP considers results from the Ames assay test to be very important in its deliberations as illustrated by the following statement from a recent Report on Carcinogens:7

DNA reactivity combined with Salmonella mutagenicity is highly correlated with induction of carcinogenicity in multiple species/sexes of rodents and at multiple tissue sites.8 A positive response in the Salmonella test was shown to be the most predictive in vitro indicator for rodent carcinogenicity (89% of the Salmonella mutagens are rodent carcinogens).9,10 Additionally, no battery of tests that included the Salmonella test improved the predictivity of the Salmonella test alone…

To eliminate the introduction of selection bias into this analysis, all positive Ames assay Salmonella bacterial mutagenicity test results reported in the literature were accepted at face value. NTP’s categorization of neoplastic evidence as either “positive” or “clear” was used to determine the tumorigenicity of the tested chemicals.

**Methods**

**Determination of neoplasticity categories 1–48**

NTP classifies the level of evidence for neoplasia as Clear (Positive), Some, Equivocal, and Negative. Analysis of the entire NTP database demonstrated that only neoplastic evidence that rose to the level of “clear” was sufficiently robust to facilitate meaningful statistical analysis.11,12 Each of the 470 chemicals for which final technical reports were available reported results for male rats, female rats, male mice, and female mice. In several cases, one of the four studies on a particular sex/species category was deemed as “inadequate” due to technical problems with that arm of the study, while the three other arms reported valid results. This situation was amenable to statistical analysis with “inadequate” ranked just higher than “negative” due to the possibility that if that arm had been completed without technical difficulty, it might have shown a level of neoplasticity higher than “negative.” The descending order of categorical rank was as follows: Clear Evidence > Some Evidence > Equivocal Evidence > Inadequate Evidence > Negative Evidence. This ranking scheme resulted in a highest category of Clear (male rats), Clear (female rats), Clear (male mice), and Clear (female mice), and a lowest category of Negative (male rats), Negative (female rats), Negative (male mice), and Negative (female mice). Due to a sporadic presentation of species/sex categories ranked as “inadequate,” the final number of categories is not set at 48 as the size of the NTP database grows, but rather that is the number of categories that result given the outcomes from the 470 current chemicals for which there are final technical reports. Online Appendix 1 shows the various combinations of Clear, Some, Equivocal, Inadequate, Negative, and the resultant categorical ranks. Figure 1 shows the number of chemicals tested per tumor potency category and Figure 2 shows the number of chemicals tested per tumor potency category (reverse order).

**Determination of ordinal rank numbers 1–135**

Analysis of the entire NTP database across all routes of administration consistently showed that the highest hurdle of neoplastic evidence was tumor site concordance across species.11,12 This result created a boundary condition under which ordinal rank could be further split within neoplasticity category (1–48), but a chemical in a lower category could not be assigned a higher ordinal rank than that of any chemical in a higher category. The second highest hurdle of neoplastic evidence was tumor site concordance across sex within species. The final criterion influencing ordinal rank was multiplicity of tumors that were not concordant by organ site. These non-concordant tumors are referred to in the ranking scheme as “single tumors.” Online Appendix 2 shows the ordinal ranking for each of the 470 chemicals resulting from simultaneous consideration of number of different tumors concordant by tumor site across species;
number of different tumors concordant across sex within species; and number of discordant tumors. Figure 3 shows the number of chemicals tested per tumor potency ordinal and Figure 4 shows the number of chemicals tested per tumor potency ordinal (reverse order).

**Determination of tumorigenicity percentile rank**

NTP currently classifies the overall level of neoplastic evidence for a particular chemical only qualitatively using the categories “Known to Be a Human Carcinogen” and “Reasonably Anticipated to Be a Human Carcinogen.” The breadth of these qualitative categories does not provide an indication of relative ranking as per tumorigenicity of the 470 chemicals tested to date for which interpretable final reports are extant. By defining the chemical with the highest ordinal rank as either 100% or 0%, a percentile rank of tumorigenicity can be assigned to any of the 470 chemicals tested to date or to any new chemical for which 2-year NTP test data are reported. In addition, each chemical can be assigned within either a quartile or quintile of tumorigenicity. Online Appendix 3 shows the percentile ranking of all 470 chemical compounds based on ordinal ranking with 2,3-dibromo-1-propanol (CASRN 96-13-9) being defined as either 100% (quintile 5) or 0% (quintile 1) since this compound has the highest tumorigenicity score via ordinal ranking.
Calculation of molecular parameters

Bio-Loom (version 1.6; Biobyte Corp., Claremont, CA, USA)\textsuperscript{12} was used to compute the three parameters used in our QSAR analysis from the simplified molecular input line entry system representation of each chemical compound: ClogP, CMR, and MgVol (Online Appendix 4). The utility of Bio-Loom for comparative QSAR (C-QSAR) analysis in comparative correlation analysis has been discussed in Hansch and Leo.\textsuperscript{5} The parameters used in this study are also discussed in detail in Hansch and Leo.\textsuperscript{5} In brief, ClogP is the calculated logarithm of the partition coefficient in octanol/water and is a measure of hydrophobicity (or lipophilicity) of a chemical.\textsuperscript{2,5} MgVol is the molar volume calculated by the method of Abraham and McGowan\textsuperscript{1,4} and CMR is the calculated molar refractivity (MR) for the whole molecule. MR is calculated as follows:

\[ MR = \left[ \frac{(n^2 - 1)}{(n^2 + 2)} \right] \times [MW / d] \]

where \( n \) is the refractive index, \( MW \) is the molecular weight, and \( d \) is the density of a substance. Since there is very little variation in \( n,^6 \) MR is largely a measure of volume with a small correction for polarizability. The MR values are scaled by 0.1. MR can be used for a substituent or for the whole molecule. ClogP and CMR are for the neutral form of partially ionized chemicals. CMR values obtained are calculated using the same program as that used to calculate ClogP.\textsuperscript{12} Note that the ClogP values are for the neutral form of acids and bases that may be partially ionized. If the degree of ionization is about the same for a set of congeners, the ionization factor can be neglected; otherwise, good correlation can be obtained using electronic terms.\textsuperscript{5,6} The correlation between experimental LogP and ClogP values for 13,815 chemicals in the CLOG program, which is a part of Bio-Loom,\textsuperscript{12} is 0.98 (experimental LogP = 1.00 ClogP − 0.03 (\( n = 13,815, r = 0.98, s = 0.35 \)). ClogP parameter that was used in this study has been widely used and cited by the QSAR community, both for environmental studies and for drug design.\textsuperscript{13–24} A very high correlation (\( r = 0.98 \)) between experimental LogP and ClogP gives confidence in using ClogP values whenever experimental LogP values are not available.

Statistical methods

The following tests were applied to assess the statistical significance of the differences in proportions.\textsuperscript{25}

Pooled test.

The null hypothesis is

\[ H_0 : p_1 - p_2 = 0 \]

The formula for the pooled test statistic comparing two proportions is

\[ z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \]

where \( \hat{p}_1 \) is the proportion in the first sample with the characteristic of interest, \( \hat{p}_2 \) the proportion in the second sample with the characteristic of interest, \( \hat{p} \) the proportion in the combined sample (all the individuals in the first and second samples together) with the characteristic of interest, and \( z \) a value on the Z-distribution.

\[ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \]

The standard error is

\[ \sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \]

Unpooled test

The null hypothesis is

\[ H_0 : p_1 - p_2 = 0 \]

\[ z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \]

Chi-squared statistic. The chi-squared (\( \chi^2 \)) statistic is defined as the sum of the squares of the Z squared values. If there are \( d \) degrees of freedom, then let this process of calculating \( \chi^2 \) continue until \( d \) different Z values are selected from the distribution. If \( Z_1, \ldots, Z_k \) are independent, standard normal random variables, then the sum of their squares,

\[ Q = \sum_{i=1}^{k} Z_i^2 \]

is distributed according to the \( \chi^2 \) distribution with \( k \) degrees of freedom. This is usually denoted as

\[ Q \sim \chi^2(k) \text{ or } Q \sim \chi_k^2 \]

The \( \chi^2 \) distribution has one parameter: \( k \), a positive integer that specifies the number of degrees of freedom (i.e. the number of \( Z_i \)'s).\textsuperscript{26}

Pearson correlation statistic. The Pearson correlation coefficient is a measure of the strength of the linear relationship between two interval or numeric variables. Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in statistics is the Pearson correlation. This correlation shows the linear relationship between two sets of data.

\[ \varphi = \sqrt{\frac{\chi^2}{n}} \]

The Pearson correlation coefficient, often referred to as the Pearson \( R \) test, is a statistical formula that measures the
strength between variables and relationships. To determine the strength of the relationship between two variables, finding the coefficient value is required, which can range between -1.00 and 1.00.

**Mann–Whitney–Wilcoxon statistic.** Generally, hypothesis testing uses techniques for testing the equality of means in two independent samples. An underlying assumption for appropriate use of the tests described was the presence of sufficiently large samples (usually \( n_1 \geq 30 \) and \( n_2 \geq 30 \)) to justify their use based on the Central Limit Theorem. For comparing two independent samples when the outcome is not normally distributed and the samples are small, a nonparametric test is appropriate.

The Mann–Whitney–Wilcoxon test is a nonparametric test to compare outcomes between two independent groups. The Mann–Whitney–Wilcoxon test is used to test whether two samples are likely to be derived from the same population (i.e. that the two populations have the same shape). Some interpret this test as comparing the medians between the two populations. A parametric test compares the means (\( H_0: \mu_1 = \mu_2 \)) between independent groups. In contrast, the null and two-sided research hypotheses for the nonparametric test are stated as follows:

\[
H_0: \text{The two populations are equal versus} \\
H_1: \text{The two populations are not equal.}
\]

The Mann–Whitney–Wilcoxon test is often performed as a two-sided test when the populations are not equal as opposed to specifying directionality. A one-sided approach is used if interest lies in detecting a positive or negative shift in one population as compared to the other. The procedure for the test involves pooling the observations from the two samples into one combined sample, keeping track of which sample each observation comes from, and then ranking lowest to highest from 1 to \( n_1 + n_2 \), respectively.

The general assumptions are as follows:

1. All the observations from both groups are independent of each other.
2. The responses are ordinal (i.e. one can at least say, of any two observations, which is the greater).
3. Under the null hypothesis \( H_0 \), the distributions of both populations are equal.
4. Under the alternative hypothesis \( H_1 \), the distributions are not equal.

The test involves the calculation of a statistic, usually called \( U \), whose distribution under the null hypothesis is known. The test statistic termed \( U \) is the smaller of \( U_1 \) and \( U_2 \), as defined in the following:

\[
U_1 = n_1n_2 + \frac{n_1(n_1 + 1)}{2} - R_1
\]

\[
U_2 = n_1n_2 + \frac{n_2(n_2 + 1)}{2} - R_2
\]

where \( R_1 = \text{sum of the ranks for group 1} \) and \( R_2 = \text{sum of the ranks for group 2} \).

For any Mann–Whitney–Wilcoxon test, the theoretical range of \( U \) is from 0 (complete separation between groups, \( H_0 \) most likely false and \( H_1 \) most likely true) to \( n_1 \times n_2 \) (little evidence in support of \( H_1 \)). In every test, \( U_1 + U_2 \) is always equal to \( n_1 \times n_2 \).

The \( Z \) statistic is used to test for significance, where

\[
Z = \frac{U - n_1(n_1 + 1)/2}{\sqrt{n_1 n_2 (n_1 + n_2)/12}}
\]

**Results**

**Relationships between Ames “positive” status, Ames “negative” status, categorical rank (1–48), and ordinal rank (1–135)**

Table 1 and Figures 5 and 6 show the relationships between Ames “positive” status, Ames “negative” status, categorical rank (1–48), and ordinal rank (1–135). The Mann–Whitney–Wilcoxon rank sum test shows that the trend in Ames versus category ranking is highly significant (\( Z = -5.69; p \text{ value near 0} \)); that is, positive Ames results are strongly associated with categorical ranks of increased
tumorigenicity. The Mann–Whitney–Wilcoxon rank sum test shows that the trend in Ames versus ordinal ranking is highly significant ($Z = -5.65$; p value near 0), that is, positive Ames results are strongly associated with ordinal ranks of increased tumorigenicity.

**Relationships between structural alerts of carcinogenesis, categorical rank (1–48), and ordinal rank (1–135)**

Table 2 and Figures 7 and 8 show the relationships between structural alerts of carcinogenesis, categorical rank (1–48), and ordinal rank (1–135). The Mann–Whitney–Wilcoxon rank sum test shows that the trend in structural alerts versus category ranking is highly significant ($Z = -7.03$; p value near 0), that is, positive structural alerts results are strongly associated with categorical ranks of increased tumorigenicity. The Mann–Whitney–Wilcoxon rank sum test shows that the trend in structural alerts versus ordinal ranking is highly significant ($Z = -7.02$; p value near 0), that is, positive structural alerts results are strongly associated with ordinal ranks of increased tumorigenicity.

**Relationships between ClogP, categorical rank (1–48), and ordinal rank (1–135)**

Table 3 shows the relationships between ClogP, categorical rank (1–48), and ordinal rank (1–135). The Mann–Whitney–Wilcoxon rank sum test shows no apparent relationship between ClogP and category ranking of tumor potency. Similarly, the Mann–Whitney–Wilcoxon rank sum test shows no apparent relationship between ClogP and ordinal ranking of tumor potency.

**Relationships between CMR, categorical rank (1–48), and ordinal rank (1–135)**

Table 4 shows the relationships between CMR, categorical rank (1–48), and ordinal rank (1–135). The Mann–Whitney–Wilcoxon rank sum test shows no apparent relationship between CMR and category ranking of tumor potency. Similarly, the Mann–Whitney–Wilcoxon rank sum test shows no apparent relationship between CMR and ordinal ranking of tumor potency.

**Relationships between MgVol, categorical rank (1–48), and ordinal rank (1–135)**

Table 5 and Figures 9 and 10 show the relationship between MgVol, categorical rank (1–48), and ordinal rank (1–135). MgVol showed an average increase with category rank of tumor potency. MgVol showed an average increase with ordinal rank of tumor potency. Therefore, smaller molecular volumes were associated with higher levels of tumorigenicity.

**Relationships between Ames Salmonella mutagenicity assay results and structural alerts of carcinogenicity**

Table 6 shows the relationships between Ames Salmonella mutagenicity assay results and structural alerts of carcinogenicity. The contingency table shows that when structural alerts of carcinogenicity were present, the Ames test was positive for 127 chemicals and the Ames test was negative 155 times. The contingency table also shows that in the absence of structural alerts of carcinogenicity there were 26 chemicals that were positive in the Ames test and 164 chemicals that were negative in the Ames test.

The null hypothesis is that the Ames test status does not correlate with structural alert status. The $\chi^2$ statistic is 50.9298. The p value is near 0. This result is significant at $p < 0.01$. The apparent correlations are that when the Ames test status is positive, then usually the structural alert status will be “yes”, whereas when the structural alert status is “no”, the Ames test status is usually negative.

The Pearson correlation $|\varphi| = \sqrt{\chi^2/N} = \sqrt{50.9298/472} = 0.329$ is not near 1.0. This most common measure of degree of association does not show strong association because
Ames negative does not predict alert status at all; alert “yes” also does not predict Ames test status.

Relationships between ClogP and Ames Salmonella mutagenicity assay results
Table 7 shows the relationships between ClogP and Ames Salmonella mutagenicity assay results. The mean ClogP for Ames positive chemicals was 1.424 (154 observations). The mean ClogP for Ames negative chemicals was 2.046 (325 observations). The difference between the ClogP means for Ames negative and Ames positive chemicals is statistically significant ($P(T \leq t)$ one-tail, 0.001; $P(T \leq t)$ two-tail, 0.002).

Relationships between ClogP and CMR and MgVol
Table 8 shows the relationships between ClogP and CMR and MgVol as calculated by the two-sample $t$-test assuming unequal variances. MgVol is highly correlated with CMR with a correlation coefficient of 0.941. MgVol is somewhat correlated with ClogP with a correlation coefficient of 0.279. CMR is somewhat correlated with ClogP with a correlation coefficient of 0.377.

Relationships between structural alerts of carcinogenicity and ClogP, CMR, and MgVol
Tables 9, 10, and 11 show the relationships between structural alerts of carcinogenicity and ClogP, CMR, and MgVol, respectively. Table 9 shows the relationship between structural alerts of carcinogenicity and ClogP. The mean ClogP when structural alerts are present is 2.170 (285 observations). The mean ClogP when structural alerts are absent is 1.393 (191 observations). The difference between the ClogP mean values for the presence and absence of structural alerts is highly statistically significant ($P(T \leq t)$ one-tail, 0.000; $P(T \leq t)$ two-tail, 0.001).

Table 10 shows the relationship between structural alerts of carcinogenicity and CMR. The mean CMR when structural alerts are present is 5.552 (281 observations). The mean CMR when structural alerts are absent is 5.192 (187 observations). The difference between the CMR mean values for the presence and absence of structural alerts is not statistically significant ($P(T \leq t)$ one-tail, 0.131; $P(T \leq t)$ two-tail, 0.261).

Table 11 shows the relationship between structural alerts of carcinogenicity and MgVol. The mean MgVol when structural alerts are present is 1.512 (285 observations). The mean MgVol when structural alerts are absent is 1.523 (191 observations). The difference between the MgVol mean values for the presence and absence of structural alerts is not statistically significant ($P(T \leq t)$ one-tail, 0.453; $P(T \leq t)$ two-tail, 0.905).

Table 12 shows a correlation matrix that summarizes the relationships noted in the text.

---

Table 2. Relationships between structural alerts of carcinogenesis, categorical rank (1–48), and ordinal rank (1–135).

| Mann–Whitney–Wilcoxon U test | Category/ordinal rank nomenclature | Ames/structural alert nomenclature | Z   | p-Value | Comment                                      |
|-----------------------------|----------------------------------|----------------------------------|-----|---------|----------------------------------------------|
| Rank sum structural alerts  | Category = 1–48                  | Structural alert                  | -7.03 | Near 0  | The rank sum test shows that the trend in structural alert versus category ranking is highly significant |
| by category ranking        |                                   | Present = 1, Absent = 0          |     |         |                                              |
| Rank sum structural alerts  | Ordinal = 1–135                   | Structural alert                  | -7.02 | Near 0  | The rank sum test shows that the trend in structural alert versus ordinal ranking is highly significant |
| by ordinal ranking         |                                   | Present = 1, Absent = 0          |     |         |                                              |

Figure 7. Relationships between structural alerts of carcinogenesis and categorical rank (1–48).

Figure 8. Relationships between structural alerts of carcinogenesis and ordinal rank (1–135).
### Table 3. Relationships between ClogP, Categorical Rank (1–48), and Ordinal Rank (1–135).

| Mann–Whitney–Wilcoxon U Test | Category / Ordinal Rank Nomenclature | Ames / Structural Alert Nomenclature | Z    | p-Value | Comment                                                                 |
|-------------------------------|--------------------------------------|-------------------------------------|------|---------|-------------------------------------------------------------------------|
| Rank Sum ClogP by Category    | Category = 1–48 Structural Alert Present = 1, Absent = 0 | Not Tested                        | Not Tested | No apparent relationship between ClogP and tumor potency (category or ordinal) |
| Rank Sum ClogP by Ordinal Ranking | Ordinal = 1–135 Structural Alert Present = 1, Absent = 0 | Not Tested                        | Not Tested | No apparent relationship between ClogP and tumor potency (category or ordinal) |

ClogP: calculated base 10 logarithm of the octanol–water partition coefficient.

### Table 4. Relationships between CMR, Categorical Rank (1–48), and Ordinal Rank (1–135).

| Mann–Whitney–Wilcoxon U Test | Category / Ordinal Rank Nomenclature | Ames / Structural Alert Nomenclature | Z    | p-Value | Comment                                                                 |
|-------------------------------|--------------------------------------|-------------------------------------|------|---------|-------------------------------------------------------------------------|
| Rank Sum CMR by Category Ranking | Category = 1–48 Structural Alert Present = 1, Absent = 0 | Not Tested                        | Not Tested | No apparent relationship between CMR and tumor potency (category or ordinal) |
| Rank Sum CMR by Ordinal Ranking | Ordinal = 1–135 Structural Alert Present = 1, Absent = 0 | Not Tested                        | Not Tested | No apparent relationship between CMR and tumor potency (category or ordinal) |

CMR: calculated molar refractivity.

### Table 5. Relationships between MgVol, Categorical Rank (1–48), and Ordinal Rank (1–135).

| Mann–Whitney–Wilcoxon U Test | Category/orordinal rank nomenclature | Ames/structural alert nomenclature | Z    | p-Value | Comment                                                                 |
|-------------------------------|--------------------------------------|-----------------------------------|------|---------|-------------------------------------------------------------------------|
| Rank sum MgVol by category ranking | Category = 1–48 Structural Alert Present = 1, Absent = 0 | Not Tested                        | Not Tested | MgVol showed an average increase with category or ordinal               |
| Rank Sum MgVol by ordinal ranking | Ordinal = 1–135 Structural Alert Present = 1, Absent = 0 | Not Tested                        | Not Tested | MgVol showed an average increase with category or ordinal               |

MgVol: McGowan molecular volume.

**Figure 9.** Relationships between MgVol and categorical rank (1–48).

**Figure 10.** Relationships between MgVol and ordinal rank (1–135).
Relationships between MgVol, Ames results, and categorical ranking of carcinogenicity (1–48)

The correlation between carcinogenicity and the combination of Ames test/MgVol can be used to improve the correlation coefficient as both variables appear to be correlated with carcinogenicity ranking.

Linear correlations were calculated for [Carcinogenicity, Ames Positive, Average MgVol], [Carcinogenicity, Ames Positive], and [Carcinogenicity, Average MgVol]. Adjusted $R^2$ are intended to compare the goodness of fit to a linear model for different choices of independent variables. The
adjusted $R^2$ values showed that only a small improvement was obtained when going from [Carcinogenicity, Ames Positive] to [Carcinogenicity, Ames Positive, Average MgVol] (i.e. [Carcinogenicity, Ames Positive] $R^2 = 0.73$ to [Carcinogenicity, Ames Positive, Average MgVol] $R^2 = 0.74$), Tables 13 and 14. Higher adjusted $R^2$ values can be obtained but their utility seems meaningless. For example, an adjusted $R^2$ value of 0.89 can be achieved if Ames Positive results, Average MgVol divided by Ames Positive values, and Average MgVol divided by Ames Negative values are used to predict the carcinogenetic potential (Table 15).

**Discussion**

The current system employed by NTP for the categorization of the neoplasticity of chemicals is qualitative. 27 Part
of the qualitative nature of the NTP categorization process is intrinsic and is due to at least two factors: (1) the less than exact nature of pathological diagnosis of pre-neoplastic and neoplastic lesions27 and (2) the practical inability to use an extremely large number of rats and mice for the purpose of increasing the statistical power of pathological observations. While these two factors necessarily introduce a qualitative aspect into the categorization of the neoplasticity observed in 2-year rodent bioassays, the large number of chemicals tested to date for which interpretable final reports are extant, that is, 470, facilitates the ability to rank these 470 chemicals and future chemical results relative to one another.

There are three different but interrelated methods for ranking these chemicals. First, neoplasticity results can be categorized from 1 to 48 at the present time by considering the various combinations of the four levels of neoplastic evidence in the descending order of categorical rank: Clear Evidence > Some Evidence > Equivocal Evidence > Inadequate Evidence > Negative Evidence (Online Appendix 1). Second, an ordinal rank 1–135 can be determined using a boundary condition under which ordinal rank can be further split within neoplasticity category (1–48), but a chemical in a lower category cannot be assigned a higher ordinal rank than that of any chemical in a higher category. When tumor site concordance across sex within species, multiplicity of tumors not concordant by organ site, and non-concordant tumors referred to in the ranking scheme as “single tumors” are considered in descending order as described in the “Methods” section and shown in Online Appendix 2, an ordinal rank number 1–135 can be readily assigned. Finally, if the most tumorigenic chemical of the 470 test results to date is defined as either 100% or 0%, a percentile ranking of each chemical ever tested or to be tested in the future logically follows (see the “Methods” section and Online Appendix 3).

The internal correlation of the categorical and ordinal ranking systems with various measures of biological activity or molecular parameters showed the expected results. The expected association of positive Ames test results with categorical and ordinal ranks of increased tumorigenicity is displayed in Table 1.28 Similarly, Table 2 shows that positive structural alerts results are strongly associated with categorical and ordinal ranks of increased tumorigenicity.29–31 Also, Table 5 demonstrates that smaller molecular volumes were associated with higher

### Table 14. Relationships between carcinogenetic potential and positive Ames Salmonella mutagenicity assay results.

| Summary output |  |
|----------------|---|
| Regression statistics |  |
| Multiple R | 0.88 |
| $R^2$ | 0.77 |
| Adjusted $R^2$ | 0.73 |
| Standard error | 10.04 |
| Observations | 7 |

| ANOVA |  |
|-------|---|
| df | SS | MS | F | Significance F |
| Regression | 1 | 1723.226 | 1723.226 | 17.0862 | 0.009053689 |
| Residual | 5 | 504.2742 | 100.8548 | | |
| Total | 6 | 2227.5 | | | |

| Coefficients | Standard error | t-Stat | p-Value | Lower 95% | Upper 95% | Lower 95.0% |
|--------------|----------------|--------|---------|-----------|-----------|-------------|
| Intercept | 66.890 | 10.483 | 6.381 | 0.001 | 39.944 | 93.837 | 39.944 |
| Ames positive | -119.296 | 28.860 | -4.134 | 0.009 | -193.484 | -45.108 | -193.484 |

| Residual output |  |
|-----------------|---|
| Observation | Predicted potency | Residuals |
| 1 | -3.49 | 4.49 |
| 2 | 9.63 | -3.63 |
| 3 | 33.49 | -17.49 |
| 4 | 31.10 | -3.10 |
| 5 | 34.68 | 5.32 |
| 6 | 35.87 | 10.63 |
| 7 | 44.22 | 3.78 |
levels of tumorigenicity as determined by categorical and ordinal ranks. 29–31

Table 7 shows the relationships between ClogP and Ames Salmonella mutagenicity assay results. The mean ClogP for Ames positive chemicals was 1.424 (154 observations). The mean ClogP for Ames negative chemicals was 2.046 (325 observations). The difference between the ClogP means for Ames negative and Ames positive chemicals is statistically significant ($P_{(T/C0)}$ one-tail, 0.001; $P_{(T/C0)}$ two-tail, 0.002).

There could be several other possible explanations for why the mean ClogP was lower for Ames positive chemicals than for Ames negative chemicals. First, the result might be artifactual since the criterion for determining whether a chemical was positive was based on whether a single positive Ames test result had been reported. Although a possibility, the large number of observations, that is, 154 observations for Ames positive chemicals and 325 observations for Ames negative chemicals, suggest that is probably not the case. Second, the collinearity between molecular size and lipophilicity might be confounding the relationship between ClogP and Ames. Specifically, as the number of hydrophobic groups on a molecule increases, the molecular size of the molecule increases. As discussed previously, smaller molecular size is associated with increased tumorigenicity (Table 5), and positive Ames test results are associated with increased tumorigenicity (Table 1). Third, both the mean ClogP value for positive Ames (1.424) and for negative Ames (2.046) represent significantly more solubility in lipid than in water, 26.55 times and 111.17 times more soluble in lipid than water, respectively.

Studies on several classes of chemicals have established that mutagenicity can be correlated with lipophilicity in a linear, parabolic, and bilinear fashion, depending upon the type of chemical class. 32–37 A

### Table 15.

| Potency | Average MgVol/ Ames positive | Average MgVol/ Ames negative | Summary output |
|---------|-----------------------------|-------------------------------|----------------|
| 1       | 0.59                        | 1.2                           | 1.3            |
| 6       | 0.48                        | 1.32                          | 1.33           |
| 16      | 0.28                        | 1.34                          | 1.51           |
| 28      | 0.3                         | 1.3                           | 1.42           |
| 40      | 0.27                        | 1.27                          | 2.05           |
| 46.5    | 0.26                        | 1.18                          | 1.68           |
| 48      | 0.19                        | 1.31                          | 1.86           |

### Regression statistics

| Regression | df | SS    | MS   | $F$   | Significance F |
|------------|----|-------|------|------|----------------|
| Regression | 3  | 2102.37 | 700.79 | 16.80 | 0.02           |
| Residual   | 3  | 125.13 | 41.71 |      |                |
| Total      | 6  | 2227.50 |       |      |                |

### Coefficients

| Coefficients | Standard error | t-Stat | p-Value | Lower 95% | Upper 95% | Lower 95.0% |
|--------------|----------------|--------|---------|-----------|-----------|--------------|
| Intercept    | 174.0          | 77.8   | 2.2     | 0.1       | –73.5     | –73.5        |
| Ames positive| –108.4         | 30.2   | –3.6    | 0.0       | –204.4    | –204.4       |
| Average MgVol/ Ames positive | –107.4 | 47.2 | –2.3 | 0.1 | –257.5 | –257.5 |
| Average MgVol/ Ames negative   | 16.3           | 14.6   | 1.1     | 0.3       | 62.7      | –30.0        |

### Residual output

| Observation | Predicted potency | Residuals |
|-------------|-------------------|-----------|
| 1           | 2.4               | –1.4      |
| 2           | 2.0               | 4.0       |
| 3           | 24.4              | –8.4      |
| 4           | 25.1              | 2.9       |
| 5           | 41.9              | –1.9      |
| 6           | 46.6              | –0.1      |
| 7           | 43.1              | 4.9       |

MgVol: McGowan molecular volume.
parabolic dependence on lipophilicity indicates that the measured biological activity of a chemical first increases with increasing lipophilicity up to an optimum value and then decreases with increasing lipophilicity. It is possible that only a chemical that is sufficiently lipophilic would be able to cross the cellular membranes and facilitate molecular transfer and thus increase tumorigenicity. Many QSAR studies for predicting mutagenicity and carcinogenicity have highlighted how individual chemicals within classes may have specific mechanisms of action. Considering the different classes of chemicals and a vast number of chemicals (470) in the data set, the ClogP correlation with tumorigenicity in this study awaits a definitive explanation.

This study represents the fourth in a series of evaluations of the entire NTP database of 594 studies, 470 of which resulted in final reports. The sequential analyses reviewed 60 inhalation studies, 212 feed studies, 124 studies by gavage, 21 via drinking water, 18 by dermal administration, and 11 by intraperitoneal injection. Across the various routes of administration, the predictive power of a positive Ames test result predicting the development of tumors in male rats, female rats, male mice, or female mice was low at approximately 35%. Similarly, the predictive power of a negative Ames test result was also low across the various routes of administration at approximately 24%. Across the various routes of administration, the predictive power of positive Ames test results predicting the development of tumors from ubiquitously neoplastic chemicals in male rats, female rats, male mice, and female mice was very low at approximately 8.3%. Similarly, the predictive power of negative Ames test results predicting the development of tumors from ubiquitously neoplastic chemicals in male rats, female rats, male mice, and female mice was also very low across the various routes of administration at approximately 5.6%. The heterogeneity of the historical database of tests of genetic toxicity other than Ames renders precise statistical analysis of this metric problematic, that is, many different tests results are reported including results from older tests, for example, sister chromatid exchange, more modern tests, for example, chromosome aberration, and less commonly conducted tests.

Conclusions
A statistical analysis of the results from the entire NTP 2-year rodent carcinogenicity database suggests two readily implementable areas of improvement. First, reliance on historical tests of genotoxicity can cloud rather than clarify the issue. It would be more cost-effective and much more definitive for the interested party (usually the manufacturer of the chemical under review) to provide a highly purified sample of the test chemical documented by a certificate of analysis to a contract laboratory previously approved by NTP and United States Environmental Protection Agency (USEPA) for the purpose of conducting a genotoxicity test battery under Good Laboratory Practices (GLP) and employing the Organization for Economic Cooperation and Development (OECD) protocol relevant to the physicochemical properties of the compound. This result would be considered the definitive evaluation of the genotoxicity of the chemical compound in question. Second, following the completion of each new NTP 2-year study, the newly tested chemical should be assigned a tumorigenicity percentile rank prior to the expert panel evaluation of the potential hazards of the chemical. In this manner, the panels would be able to provide a relative perspective on the potential carcinogenicity of the chemical.

These suggestions for improvement may seem idealistic in the current environment of toxicity testing and may not be implementable at this time due to the variety of methodologies used, inter-lab variations, reporting and evaluation differences, purity of substances tested, etc. However, in light of the present situation, efforts must be made to improve the testing and reporting methods currently in place.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental material
Supplementary material for this article is available online.

References
1. National Toxicology Program. NTP vision & roadmap future directions. https://ntp.niehs.nih.gov/about/vision/index.html (2016, accessed 15 December 2016).
2. Leo A. Calculating log P_{oct} from structures. Chem Rev 1993; 93: 1281–1306.
3. Abraham MH. Scales of solute hydrogen-bonding—their construction and application to physicochemical and biochemical processes. Chem Soc Rev 1993; 22: 73–83.
4. Abraham MH and McGowan JC. The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography. Chromatographia 1987; 23: 243–246.
5. Hansch C and Leo A. Exploring QSAR: fundamentals and applications in chemistry and biology. Washington, DC: American Chemical Society, 1995.
6. Hansch C, Leo A and Hoekman D. Exploring QSAR: hydrophobic, electronic, and steric constants. Washington, DC: American Chemical Society, 1995.
7. National Toxicology Program. US Department of Health and Human Services. Report on carcinogens, monograph on 1-bromopropane. September 2013, p. 36. https://ntp.niehs.nih.gov/ntp/roc/thirteenth/monographs_final/1bromopropane_508.pdf (accessed 1 November 2016).
8. Ashby J and Tennant RW. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the US NTP. Mutat Res 1991; 257: 229–306.
9. Agency for Toxic Substances and Disease Registry (ATSDR). Toxicological profile of antimony and related compounds. September 1992. http://www.atsdr.cdc.gov/toxprofiles/tp23.pdf (1992, accessed 15 December 2016).

10. Tennant RW, Margolin BH, Shelby MD, et al. Prediction of chemical carcinogenicity in rodents from in vitro genetic toxicity assays. Science 1987; 236: 933–941.

11. Smith CJ and Perfetti TA. Tumor site concordance and genetic toxicology test correlations in NTP 2-year feed studies. Toxicol Res Appl 2017; 1: 1–12.

12. BioByte Corp. Biobyte—Bio-Loom. 2016. http://biobyte.com/bb/prod/bioloom.html (accessed 9 February 2017).

13. Mannhold R, Poda GI, Ostermann C, et al. Calculation of molecular lipophilicity: state-of-the-art and comparison of log P methods on more than 96,000 compounds. J Pharm Sci 2009; 98: 861–893.

14. Arnot JA and Gobas FAPC. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environ Res 2006; 14: 257–297.

15. Devillers J, Domine D, Bintein S, et al. Comparison of fish bioconcentration models. In: Devillers J (ed.) Comparative QSAR. Washington, DC: Taylor & Francis, 1998, pp. 1–50.

16. Garg R, Gupta SP, Gao H, et al. Comparative quantitative structure–activity relationship studies on anti-HIV drugs. Chem Rev 1999; 99: 3525–3602.

17. Hansch C, Kim D, Leo AJ, et al. Toward a quantitative comparative toxicology of organic compounds. Crit Rev Toxicol 1989; 19: 185–226.

18. Leo A and Hansch C. Role of hydrophobic effects in mechanistic QSAR. Pers Drug Des Discov 1999; 17: 1–25.

19. Muller M and Nendza M. Literature study: Comparative analysis of estimated and measured BCF data (OECD 305) with a special focus on differential accumulation of (mixtures of) stereoisomers, 2009, Dessau-Roßlau, Germany: Federal Environment Agency (Umweltbundesamt), http://www.uba.de/uba-info-median-e-4088.html (accessed 9 February 2017).

20. Selassie CD, Garg R and Mekapati SB. A mechanism-based approach to the study of the toxicity of endocrine disruptive agents. Pure Appl Chem 2003; 75: 2363–2374.

21. Smith CJ, Perfetti TA, Morton MJ, et al. The relative toxicity of substituted phenols reported in cigarette mainstream smoke. Tox Sci 2002; 69: 265–278.

22. Smith CJ, Perfetti TA, Garg R, et al. IARC carcinogens reported in cigarette mainstream smoke and their calculated log P values. Food Chem Toxicol 2003; 41: 807–817.

23. Smith CJ, Perfetti TA, Garg R, et al. Percutaneous penetration enhancers in cigarette mainstream smoke. Food Chem Toxicol 2004; 42: 9–15.

24. Smith CJ, Perfetti TA, Garg R, et al. Utility of the mouse dermal promotion assay in comparing the tumorigenic potential of cigarette mainstream smoke. Food Chem Toxicol 2006; 44: 1699–1706.

25. Motulsky H. Intuitive biostatistics: a nonmathematical guide to statistical thinking, 2nd ed. New York: Oxford University Press, 2010.

26. Agresti A. An introduction to categorical data analysis. New York: Wiley, 1996.

27. EPA (U.S. Environmental Protection Agency). Guidelines for carcinogen risk assessment. Federal Register 2005; 70(66): 17766–17817. https://www.gpo.gov/fdsys/pkg/FR-2005-04-07/pdf/05-6642.pdf (accessed 7 April 2005).

28. Benigni R and Bossa C. Alternative strategies for carcinogenicity assessment: an efficient and simplified approach based on in vitro mutagenicity and cell transformation assays. Mutagenesis 2011; 26(3): 455–460.

29. Benigni R and Bossa C. Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational database technology. Mutat Res 2008; 659: 248–261.

30. Benigni R, Bossa C and Tcheremenskaia O. Nongenotoxic carcinogenicity of chemicals: mechanisms of action and early recognition through a new set of structural alerts. Chem Rev 2013; 133: 2940–2957.

31. Plošnik A, Vračko M and Sollner Dolenc M. Mutagenic and carcinogenic structural alerts and their mechanisms of action. Arch Hig Rada Toksikol 2016; 67: 169–182.

32. Drug discovery and evaluation: Safety and pharmacokinetic assays. In: Vogel HG, Maas J, Hock FJ and Mayer D (eds) Silicio methods. Berlin, Heidelberg, New York: Springer, 2006, p. 804.

33. Lopez De Compadre RL, Shusterman AJ and Hansch C. The role of hydrophobicity in the Ames test. The correlation of the mutagenicity of nitropoly cyclic hydrocarbons with partition coefficients and molecular orbital indices. Int J Quant Chem 1988; 34(2): 91–101.

34. National Research Council (US) Steering Committee on Identification of Toxic and Potentially Toxic Chemicals for Consideration by the National Toxicology Program. Appendix D: the analysis of structure-activity relationships in selecting potentially toxic compounds for testing. In: Norman Grossblatt (ed.) Toxicity testing: strategies to determine needs and priorities. Washington, DC: National Academies Press, 1984.

35. Smith CJ and Perfetti TA. Tumor site concordance and genetic toxicology test correlations in NTP two-year gavage, drinking water, dermal, and intraperitoneal injection studies. Toxicol Res Appl. 2018. doi: 10.1177/2397847317751147.

36. Benigni R, Passerini L, Gallo G, et al. QSAR models for discriminating between mutagenic and nonmutagenic aromatic and heteroaromatic amines. Environ Mol Mutagen 1998; 2: 75–83.

37. Debnath AK, Lopez de Compadre RL, Shusterman AJ, et al. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 2. Mutagenicity of aromatic and heteroaromatic nitro chemicals in Salmonella typhimurium TA100. Environ Mol Mutagen 1992; 19: 53–70.

38. Smith CJ and Anderson SP. High discordance in development and organ site distribution of tumors in rats and mice in NTP 2-year inhalation studies. Toxicol Res Appl 2017; 1: 12–22.