J. Fac. Pharm. Ankara / Ankara Ecz. Fak. Derg., 48(1): 20-33, 2024
Doi: 10.33483/jfpau.1322948

ORIGINAL ARTICLE / ÖZGÜN MAKALE

A MULTIVARIATE INTERPOLATION APPROACH FOR PREDICTING DRUG LD_{50} VALUE

İLAÇ LD_{50} DEĞERİNİ TAHMİN ETMEK İÇİN ÇOK DEĞİŞKENLİ BİR İNTERPOLASYON YAKLAŞIMI

Gül KARADUMAN[1,2]*, Feyza KELLECİ ÇELİK[1]*

[1]Karamanoğlu Mehmetbey University, Vocational School of Health Services, 70200, Karaman, Türkiye
[2]University of Texas at Arlington, Department of Mathematics, TX 76019-0408, Arlington, USA

ABSTRACT

Objective: *The present study aimed to develop a multivariate interpolation based on the quantitative structure-toxicity relationship (QSTR) that can accurately predict the oral median lethal dose (LD_{50}) values of drugs in mice by considering five different toxicologic endpoints.*
Material and Method: *A mathematical model was created using a comprehensive dataset comprising LD_{50} values from 319 pharmaceuticals belonging to various pharmacological classes. We developed a polynomial model that can predict the range of LD_{50} values for pharmaceuticals. We employed a technique called two-variable polynomial interpolation. This method allowed us to estimate the approximate values of a function at any point within a two-dimensional (2D) space by utilizing a polynomial equation.*
Result and Discussion: *The resulting model demonstrated the ability to predict LD_{50} values for new or untested drugs, rendering it a valuable tool in the early stages of drug development. The Ghose-Crippen-Viswanadhan octanol-water partition coefficient (ALogP) and Molecular Weight (MW) were selected as suitable descriptors for building the best QSAR model. Based on our evaluation, the model achieved an overall success rate of 86.73%. Compared to traditional experimental methods for LD_{50} determination, this innovative approach offers time and cost efficiency while reducing animal testing requirements. Our model can improve drug safety, optimize dosage regimens, and assist decision-making processes during preclinical studies and drug development. This approach provided a reliable and efficient method for preliminary acute toxicity assessments.*
Keywords: *Data analysis, LD_{50}, mathematical toxicology, multivariate interpolation, polynomial interpolation*

ÖZ

Amaç: *Bu çalışmanın amacı, beş farklı toksikolojik sonucu dikkate alarak farelerde ilaçların oral median letal doz (LD_{50}) değerlerini doğru bir şekilde tahmin edebilen, niceliksel yapı-toksisite ilişkisine (QSTR) dayalı çok değişkenli bir interpolasyon yöntemi geliştirmektir.*
Gereç ve Yöntem: *Farklı farmakolojik sınıflara ait 319 ilaca ait LD_{50} değerlerini içeren kapsamlı bir veri seti kullanılarak matematiksel bir model oluşturuldu. Farmasötiklerin LD_{50} değerlerinin aralığını tahmin edebilen bir polinom model geliştirdik. İki değişkenli polinom interpolasyon adı*

* **Corresponding Author / Sorumlu Yazar:** Gül Karaduman
e-mail / e-posta: gulk@bu.edu, **Phone / Tel.:** +903382262798

* **Corresponding Author / Sorumlu Yazar:** Feyza Kelleci Çelik
e-mail / e-posta: feyza-kelleci@hotmail.com, **Phone / Tel.:** +903382262761

Submitted / Gönderilme : 05.07.2023
Accepted / Kabul : 22.09.2023
Published / Yayınlanma : 20.01.2024

verilen bir teknik kullanarak bunu gerçekleştirdik. Bu yöntem, bir polinom denklemi kullanarak iki boyutlu bir uzayda herhangi bir noktadaki bir fonksiyonun değerlerini tahmin etmemizi sağladı.
Sonuç ve Tartışma: *Elde edilen model, yeni veya denenmemiş ilaçlar için LD_{50} değerlerini tahmin etme yeteneğini gösterdi ve bu nedenle ilaç geliştirme sürecinin erken aşamalarında değerli bir araç olarak kullanılabilir. Değerlendirmemize göre, model genel başarı oranı olarak %86,73 olarak bulundu. LD_{50} değerinin belirlenmesinde kullanılan geleneksel deneysel yöntemlere kıyasla, bu yenilikçi yaklaşım zaman ve maliyet açısından avantajlı olup hayvan deneylerinin gerekliliğini azaltmaktadır. Modelimiz ilaç güvenliğini artırabilir, doz rejimlerini optimize edebilir ve ön klinik çalışmalar ve ilaç geliştirme sürecinde karar verme süreçlerine yardımcı olabilir. Bu yaklaşım, ön akut toksisite değerlendirmeleri için güvenilir ve etkili bir yöntem sunmuştur.*
Anahtar Kelimeler: *Çok değişkenli interpolasyon, LD_{50}, matematiksel toksikoloji, polinom interpolasyonu, veri analizi*

INTRODUCTION

The median lethal dose or concentration (LD_{50}/LC_{50}) serves as a dose indicator for evaluating the acute toxicity of pharmaceuticals/chemicals in risk assessment. This dosage corresponds to the amount resulting in mortality in 50% of the analyzed animal population. The LD_{50}/LC_{50} value is also used for categorizing the toxicity levels of substances, enabling a standardized and systematic approach to toxicological assessments (Table 1) [1].

LD_{50}/LC_{50} tests are conducted in the early stages of drug development to determine the lethal dose of pharmaceuticals. These trials provide a reference point for dose selection in subsequent toxicity studies. The accurate calculation of the LD_{50}/LC_{50} value is essential for ensuring the safe use of medications and predicting potential adverse reactions. Thus, it aids in the establishment of the drug's toxicity profile [2]. Although rats, rabbits, and guinea pigs have traditionally been utilized in such studies, mice are often preferred as a model organism. The LD_{50}/LC_{50} value can be determined through various administration methods, including oral, dermal, or inhalation, depending on the study's design. Regarding ease of use, the oral route is the commonly preferred method of medication delivery. Therefore, oral LD_{50}/LC_{50} data in the literature are widely available compared to other routes of exposure [1].

Table 1. The hazard categories for acute toxicity (based on the LD_{50}/LC_{50} value) according to the Globally Harmonized System of Classification and Labelling of Chemicals (GHS) guidelines [1]

Exposure Route	Category 1	Category 2	Category 3	Category 4	Category 5
Oral (mg/kg BW)	≤ 5	5-50	50-300	300-2000	2000 <
Dermal (mg/kg BW)	≤ 50	50-200	200-1000	1000-2000	2000 <
Gases (ppmV)	≤ 100	100-500	500-2500	2500-20000	-
Vapors (mg/l)	≤ 0.5	0.5-2	2-10	10-20	-
Dust and Mists (mg/l)	≤ 0.05	0.05-0.5	0.5-1	1-5	-

BW: Body Weight

Due to ethical concerns, regulatory rules have recently restricted the use of laboratory animals in pharmaceutical research [3]. The current legislations promote the adoption of alternative approaches to minimize animal experiments [4]. One of the notable alternative methods involves the quantitative structure-toxicity relationship (QSTR) models, which assist in the rapid and cost-effective assessment of drug toxicity. QSTR modeling has gained recognition as a computational approach in pharmaceutical toxicology. These mathematical-based models establish a relationship between the structural characteristics of pharmaceutical compounds and their potential toxic effects. Various algorithms are employed in QSTR studies to facilitate classification or provide direct quantitative predictions [5]. Several studies have been conducted in the literature to predict the LD_{50}/LC_{50} value of chemical substances using QSTR models, employing datasets of varying sizes [4,6,7]. However, compared to other research areas in QSTR, relatively few studies specifically focused on predicting the LD_{50}/LC_{50}

values of pharmaceuticals [8]. Due to this gap in the literature, we have turned to non-animal-based methods to predict the LD_{50}/LC_{50} values of pharmaceuticals.

The interpolation method is one of the commonly employed approaches. If LD_{50}/LC_{50} values for certain drugs are available at specific doses, interpolation techniques can be used to estimate LD_{50}/LC_{50} values for intermediate doses [9]. Interpolation methods such as linear interpolation, polynomial interpolation, or spline interpolation can be used to approximate LD_{50}/LC_{50} values based on the known data points [4]. Another approach is the regression technique. This technique can be used to estimate LD_{50}/LC_{50} values based on a set of independent variables (predictors) such as drug dosage, administration route, or changing experimental conditions [4]. Various regression techniques like linear, logistic, or nonlinear regression can be applied to fit a model to the available LD_{50}/LC_{50} data and predict LD_{50}/LC_{50} values for newly synthesized drugs or different dosages. QSTR models also aim to establish relationships between the chemical structure or descriptors of drugs and their biological activities. There are critical dose values that play a significant role in the biological activity of a drug. Among these criteria, the LD_{50}/LC_{50} value stands out as an indicator of acute toxicity [4,6]. By analyzing a dataset of drugs with known LD_{50}/LC_{50} values and their corresponding chemical descriptors, QSTR models can be built to predict LD_{50}/LC_{50} values for novel drug molecules based on their structural characteristics. Machine learning techniques, such as decision trees, random forests, support vector machines, or neural networks, can be employed to develop predictive models for LD_{50}/LC_{50} estimation [10,11]. These models learn patterns and relationships from the available LD_{50}/LC_{50} data and can be used to predict LD_{50}/LC_{50} values for novel drugs based on their features or descriptors. The choice of the most appropriate method varies depending on the available data, the nature of the problem, and the specific goals of the analysis.

In this study, we developed a multivariate interpolation-based [12] QSTR model to predict the acute oral LD_{50} values of drugs in mice. This model was formulated based on the impact of critical properties in the chemical structures of pharmaceuticals on the biological response in mice. This approach enables the determination of relationships between the physicochemical properties of pharmaceuticals and their toxicity, allowing for the rapid and effective analysis and interpretation of complex data. This provides a significant advantage in reducing risks and improving safety standards in drug development. Our study was specifically designed to minimize the utilization of experimental animals by narrowing down the range of LD_{50} values. Ultimately, a final LD_{50} value should be established through animal experimentation in the last stage, thus ensuring comprehensive assessment and verification of drug safety. The results obtained from this research contribute to the advancement of drug safety assessment and establish a foundation for future computational toxicology studies.

Figure 1 illustrates the steps to develop a mathematical model capable of predicting the range of LD_{50} values for pharmaceuticals-the initial phase involved data collection. Subsequently, the dataset was pre-processed by eliminating noisy data and establishing the applicability domain. From the dataset, specific drugs were selected for constructing the interpolation polynomial. Two descriptors with the highest efficiency were chosen to represent the two variables, (x and y). Multiple interpolation polynomials were created to assess their success in predicting the LD_{50} value range. If the results were unsatisfactory, the process returned to step 3, where different sets of drugs were selected, and the interpolation process was repeated. Ultimately, the model's performance was evaluated based on the accuracy of correctly classifying the drug ranges of LD_{50} values.

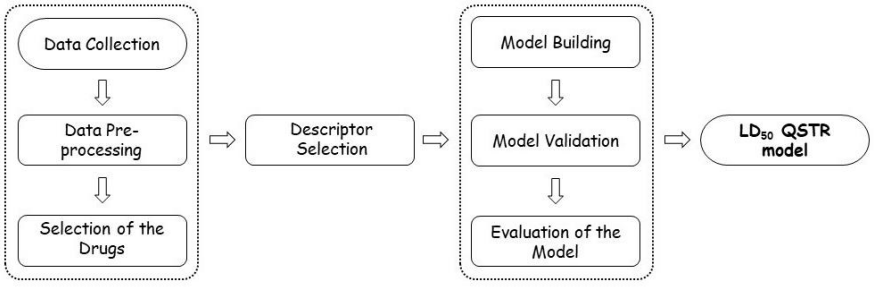

Figure 1. Model development workflow

MATERIAL AND METHOD

Polynomial interpolation, in the context of QSTR, is a mathematical technique used to model the relationship between the chemical structure of a compound and its toxicological activity. QSTR models aim to predict the activity of a chemical based on its structural features. Polynomial interpolation involves fitting a polynomial function to a set of data points, where each data point represents a compound with known structural descriptors and corresponding activity values. The polynomial function is then used to interpolate the activity of compounds based on their structural descriptors.

Estimating the LD_{50} values for drugs solely through interpolating molecular descriptors is challenging and not prevalent. Molecular descriptors alone may not provide sufficient information to predict toxicity levels accurately. Rather than predicting a direct mathematical value, this method yields more successful results in estimating a range. Therefore, in this study, we developed a mathematical model that efficiently predicts the range of LD_{50} values for a pharmaceutical.

Material

This study included a total of 319 drugs, each accompanied by available oral LD50 values [13-15] (Supplementary File 1_Table S1). To ascertain the compounds' chemical structure and physical attributes, we accessed the two-dimensional structural data file (2D SDF) through the PubChem database [16]. Employing the open-source program T.E.S.T. [17], we computed chemical descriptors for these compounds based on the 2D SDF data files.

We diligently cleansed and preprocessed the dataset, ensuring its freedom from any missing values, outliers, or other data quality anomalies. The importance of data quality cannot be understated in the realm of data science and machine learning workflows. Consequently, any corrupted 2D SDF files sourced from PubChem were meticulously eliminated from the dataset, yielding data of exceptional quality and usability.

Subsequently, the remaining SDF files underwent characterization through T.E.S.T. and were then stored in comma-separated files. The "ReplaceMissingValues" tool, an unsupervised attribute filter within the WEKA 3.9.5 (Waikato Environment for Knowledge Analysis) software, was employed to impute the missing descriptive values [23]. Following this step, duplicated data entries were systematically removed from the dataset to ensure its integrity.

The selection of descriptors for accurate LD_{50} estimation depends on the specific characteristics of the drugs being considered. We focused on the two descriptors frequently mentioned in the literature for the interpolation process. Two commonly used and important descriptors for LD_{50} estimation are the Ghose-Crippen-Viswanadhan octanol-water partition coefficient (ALogP) and Molecular Weight (MW). The selection of ALogP and MW as significant descriptors for LD_{50} estimation is based on the following reasons.

ALogP is employed to assess a compound's solubility in nonpolar solvents (e.g., octanol) and polar solvents (e.g., water) [18]. It provides information about the lipophilicity of a drug and its ability to permeate and accumulate in biological membranes. Higher ALogP values are associated with increased toxicity, potentially due to enhanced membrane permeability and the likelihood of accumulating in fatty tissues [5]. Therefore, ALogP is a valuable indicator of a compound's bioavailability and potential to cross biological membranes, making it a significant factor in determining LD_{50} values.

Our other identifier, MW, is a fundamental descriptor used in drug-related QSTR studies, providing information about a drug's size and complexity. MW is a critical parameter in toxicity modeling studies as it directly influences the toxicokinetic of xenobiotics. According to the OECD guidelines, it has been stated that the physicochemical properties of a chemical, such as AlogP and MW, may be helpful for study planning and interpretation of results [19].

Molecular Diversity and Distribution in Chemical Space

Molecular diversity and distribution in chemical space play a crucial role in QSTR modeling. Chemical space refers to the vast multidimensional space that encompasses all possible molecules. The distribution of molecules within this chemical space is important because it affects the coverage and

representativeness of the training data used in QSTR modeling. Molecular diversity measures the variety and heterogeneity of molecules in a dataset. A diverse dataset should cover different regions of chemical space, representing a wide range of structural features and properties. Including diverse molecules in the training set helps capture the full range of interactions and properties that a QSTR model needs to predict accurately [20].

Considering molecular diversity and distribution in chemical space, QSTR models can provide reliable predictions for molecules with similar structural features or properties, even if they were not explicitly included in the training dataset. This enhances the generalization and applicability of QSTR models to guide molecular design, screening, and optimization processes in various fields [20].

Method

In this study, we aimed to develop a polynomial model that can predict the range of oral LD_{50} values for pharmaceuticals. We employed a technique called two-variable polynomial interpolation. This method allows us to estimate the values of a function at any point within a 2D space by utilizing a polynomial equation.

We have a set of data points, each consisting of distinct values $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, along with their corresponding function values $f(x_i, y_i)$. Our goal was to find a unique multivariate interpolation polynomial, denoted as P(x, y), that satisfies the equation:

$$f(x_i, y_i) = P(x_i, y_i), \tag{1}$$

where, for each $i = 0,1, \dots, n$. This equation should hold true for each i ranging from 0 to n. By constructing such a polynomial, we aimed to accurately predict the function values for any given combination of x and y within the interpolation domain. To accurately predict function values for any combination of x and y within the interpolation domain, we need to create an interpolation matrix. The construction of this matrix can be accomplished using the following procedure.

The polynomial of two variables of the total degree of n is given by

$$P(x, y) = \sum_{i=0}^{n} \sum_{j=0}^{k} a_{j,i}\, x^j y^{i-j}, \tag{2}$$

where, for each $i = 0,1, \dots, n$ and $j = 0,1, \dots, k$ [21]. We used 10 distinct (x, y) values to find a multivariate interpolation polynomial function $P(x, y)$ of the form,

$$\begin{aligned} P(x, y) = {} & a_{0,1} + a_{1,1}x + a_{1,2}y + a_{2,1}x^2 + a_{2,2}xy + a_{2,3}y^2 + a_{3,1}x^3 + a_{3,2}x^2y \\ & + a_{3,3}xy^2 + a_{3,4}y^3 \end{aligned} \tag{3}$$

where $a_{0,1}, a_{1,1}, a_{1,2}, a_{2,1}, a_{2,2}, a_{2,3}, a_{3,1}, a_{3,2}, a_{3,3}$, and $a_{3,4}$ are the coefficients to be determined.

We can construct a system of equations by substituting the data points into the polynomial equation,

$$\begin{aligned} f(x_1, y_1) = {} & a_{0,1} + a_{1,1}x_1 + a_{1,2}y_1 + \cdots + a_{3,1}{x_1}^3 + a_{3,2}{x_1}^2y_1 + a_{3,3}x_1{y_1}^2 \\ & + a_{3,4}{y_1}^3 \\ f(x_2, y_2) = {} & a_{0,1} + a_{1,1}x_2 + a_{1,2}y_2 + \cdots + a_{3,1}{x_2}^3 + a_{3,2}{x_2}^2y_2 + a_{3,3}x_2{y_2}^2 \\ & + a_{3,4}{y_2}^3 \\ & \vdots \\ f(x_{10}, y_{10}) = {} & a_{0,1} + a_{1,1}x_{10} + a_{1,2}y_{10} + \cdots + a_{3,1}{x_{10}}^3 + a_{3,2}{x_{10}}^2y_{10} \\ & + a_{3,3}x_{10}{y_{10}}^2 + a_{3,4}{y_{10}}^3, \end{aligned} \tag{4}$$

where, $a_{0,1}$, $a_{1,1}, \ldots, a_{3,4}$ are the coefficient values that are to be determined to form the interpolation polynomial $P(x, y)$ [22,23]. We can represent this system of equations in matrix form,

$$A\,a\; =\; f, \tag{5}$$

where A is the coefficient matrix, a is the vector of unknown coefficients, and f is the vector of function values. Equation (5) can be expressed as a linear system,

$$Aa = \begin{bmatrix} 1 & x_1 & \cdots & x_1{y_1}^2 & {y_1}^3 \\ 1 & x_2 & & x_2{y_2}^2 & {y_2}^3 \\ \vdots & & \ddots & \vdots & \\ 1 & x_9 & \cdots & x_9{y_9}^2 & {y_9}^3 \\ 1 & x_{10} & & x_{10}{y_{10}}^2 & {y_{10}}^3 \end{bmatrix} \begin{bmatrix} a_{0,1} \\ a_{1,1} \\ \vdots \\ a_{3,3} \\ a_{3,4} \end{bmatrix} = \begin{bmatrix} f(x_1, y_1) \\ f(x_2, y_2) \\ \vdots \\ f(x_9, y_9) \\ f(x_{10}, y_{10}) \end{bmatrix} = f, \tag{6}$$

where, $A \in \mathbb{R}^{10\times10}$, $f \in \mathbb{R}^{10}$, and $a \in \mathbb{R}^{10}$. In this equation, A is a real-valued invertible matrix, f is a vector in the real-valued space, and a is the vector we need to find. Once we have matrix A and vector f, we can solve for vector $a \in \mathbb{R}^{10}$ using matrix operations or linear regression techniques to obtain the coefficients. These coefficients represent the interpolated function. We can then evaluate the interpolated function at new points within the interpolation domain by substituting the input values and obtaining the estimated function values.

The degree of the polynomial can be adjusted based on the targeted level of accuracy and complexity. Higher-degree polynomials can provide a better fit to the data but may also lead to overfitting. Careful consideration of the dataset and the trade-off between accuracy and complexity is critical when choosing the degree of the polynomial for interpolation. In our study, we chose this option since the third-degree polynomial produced the greatest results. The framework of multivariate interpolation (MVI) is shown in Figure 2.

Input: The input points x, y
Output: Corresponding target values $f(x, y)$
function coefficients = findInterpolationCoefficients(x, y, f)
% Create the matrix A
Step 1: $A = \begin{bmatrix} 1 & x_1 & \cdots & x_1{y_1}^2 & {y_1}^3 \\ 1 & x_2 & & x_2{y_2}^2 & {y_2}^3 \\ \vdots & & \ddots & \vdots & \\ 1 & x_9 & \cdots & x_9{y_9}^2 & {y_9}^3 \\ 1 & x_{10} & & x_{10}{y_{10}}^2 & {y_{10}}^3 \end{bmatrix}$;

% Create the vector f
Step 2: $f = \begin{bmatrix} f(x_1, y_1) \\ f(x_2, y_2) \\ \vdots \\ f(x_9, y_9) \\ f(x_{10}, y_{10}) \end{bmatrix}$;
% Solve the linear system $Aa = f$
Step 3: coefficients = $A \,/\, f$;
% Output: Estimated $f(x, y)$
estimated_f = $f(x, y)$;

Figure 2. Algorithm: MVI

Model Validation

Evaluating the performance of the QSTR model on the dataset is crucial to assess its predictive ability across diverse molecules. The test sets should contain molecules that are structurally distinct from the training set, representing novel regions of chemical space. The accuracy of the polynomial was evaluated using the accuracy (ACC) metric. Additionally, a visual analysis of the graph depicting a

polynomial curve is presented.

RESULT AND DISCUSSION

The performance of the multivariate interpolation model was evaluated using various metrics to assess its accuracy and effectiveness in estimating LD_{50} values based on molecular descriptors. We employed the interpolation polynomial $P(x,y)$, utilizing two molecular descriptor values, MW and AlogP, for each drug. The MW values represented the x-axis, while the AlogP values represented the y-axis for constructing the interpolation polynomial $P(x,y)$. For our analysis, we selected a dataset consisting of 10 drugs with their corresponding MW, AlogP, and LD_{50} values. We defined the output function values $f(x_i,y_i)$ based on the LD_{50} values. By inserting the interpolation points (x_i,y_i) into a system of equations, we constructed a coefficient matrix $A \in \mathbb{R}^{10\times10}$, an output vector $f \in \mathbb{R}^{10}$, and an unknown vector $a \in \mathbb{R}^{10}$. The coefficient matrix A had to be non-singular to ensure a unique solution vector $a \in \mathbb{R}^{10}$ [24]. To determine the singularity of A, we computed the number of linearly independent columns, which signifies the columns that are not linear combinations of each other. All 10 columns of A were linearly independent, indicating that matrix A was non-singular or invertible. This guaranteed that the system of equations had a single solution. To calculate the values of $a_{j,i}$, representing the coefficients of the interpolation polynomial $P(x,y)$, we utilized MATLAB, a powerful computational tool commonly used for scientific calculations and data analysis. MATLAB facilitated the efficient solution of the system of equations and enabled us to obtain the coefficients of the interpolation polynomial $P(x,y)$. The coefficient values $a_{j,i}$ for the model $P(x,y)$ with two variables were calculated based on the system (5) and are presented in Table 2.

Table 2. Coefficient values $a_{j,i}$

Coefficient	Calculated Value of Coefficient
$a_{0,1}$	-91.9388
$a_{1,1}$	385.1059
$a_{1,2}$	134.0851
$a_{2,1}$	-526.8917
$a_{2,2}$	-450.4742
$a_{2,3}$	125.3757
$a_{3,1}$	235.0836
$a_{3,2}$	355.8234
$a_{3,3}$	-84.6011
$a_{3,4}$	-165.7975

Once the $a_{j,i}$ values calculated, the interpolation polynomial $P(x,y)$ was determined as follows,

$$P(x,y) = -91.9388 + 385.1059x + 134.0851y - 526.8917x^2 - 450.4742xy + 125.3757y^2 + 235.0836x^3 + 355.8234x^2y - 84.6011xy^2 - 165.7975y^3, \quad (7)$$

where x is MW and y is AlogP descriptor values of the drugs.

Figure 3 illustrates the visual representation of the interpolation polynomial $P(x,y)$, showing the relationship between two variables, MW (Molecular Weight) and AlogP values (partition coefficient), of various drugs. The x-axis corresponds to the MW values, while the y-axis represents the AlogP values. In Figure 3, the interpolation polynomial is visualized as a curve that smoothly passes through the data points. Each data point represents a specific drug, with the MW and AlogP values corresponding to its position on the x and y-axis, respectively.

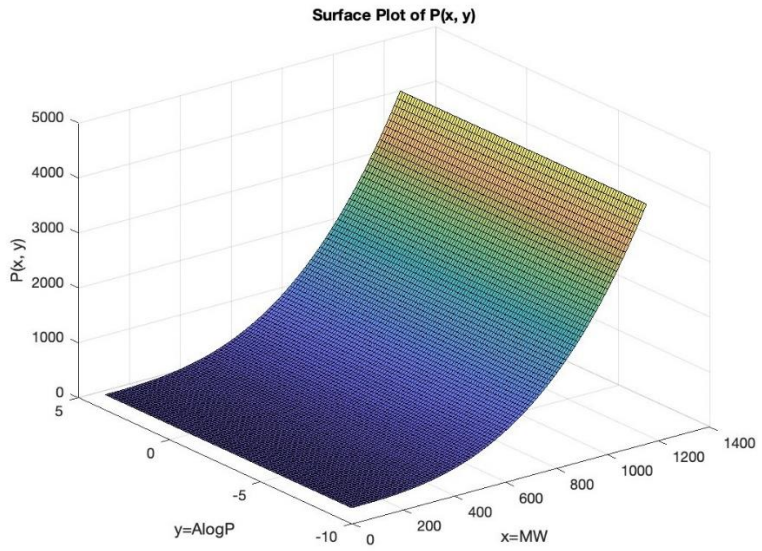

Figure 3. Surface plot of the interpolation polynomial

Model Validation

The model underwent both internal and external validations. Internal validation of a QSTR model involves assessing its accuracy using the molecules employed during its creation. This includes predicting the activities of these molecules and analyzing parameters to determine prediction precision. The accuracy of the multivariate interpolation model during internal validation was assessed by estimating LD_{50} values for the chosen 10 drugs used in creating the polynomial P(x, y). Remarkably, the internal validation achieved a 100% success rate, given that the same drugs were used for both development and testing.

However, the model's ability to predict outcomes for entirely new compounds cannot be reliably determined based solely on internal validation, as it relies on the same compounds utilized in its development. To contend with this matter, external validation becomes instrumental. In this scenario, the dataset is partitioned into training and test sets, comprising 309 and 10 drugs, respectively. The model is constructed using the training set and subsequently validated using the independent test set, ensuring its adaptability to novel compounds. The accuracy of the multivariate interpolation model for the external validation set was evaluated by estimating LD_{50} values for the selected drugs using the input MW and AlogP values. To quantify the accuracy, we employed the ACC metric, which measures the proportion of correct predictions made by the model out of the total number of predictions. Using the interpolation polynomial, we categorized the drugs based on the ranges provided in Table 2. For instance, when estimating the LD_{50} value of a drug, we applied the interpolation polynomial within the appropriate range indicated in Table 2. If the estimated LD_{50} value fell within the correct range, we considered it a correct estimate.

Conversely, if the estimated value corresponded to a different interval, it was considered a false estimate. Based on our evaluation, the model achieved an overall success rate of 86.73%. This means that out of the 309 drugs tested, we correctly predicted the category (range) for 268 drugs. The high success rate indicates the model's proficiency in accurately estimating the LD_{50} values for a significant portion of the tested drugs. Figure 4 presents a comparison between the experimental and calculated LD_{50} values of the drugs. As depicted in the graph, the results obtained through the interpolation polynomial exhibit a notable level of success.

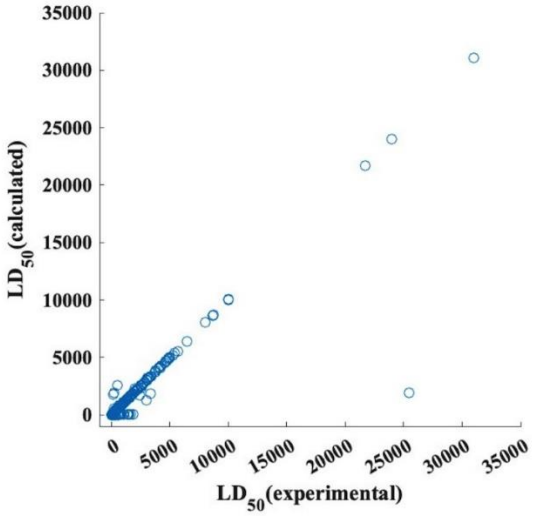

Figure 4. Experimental vs. Calculated LD_{50} values

Our analysis demonstrated that the multivariate interpolation approach yielded accurate LD_{50} predictions, with low errors suggesting a strong correlation between the estimated and actual LD_{50} values. This finding emphasizes the effectiveness of the multivariate interpolation model in estimating acute drug toxicity levels based on the molecular descriptors MW and AlogP. The ACC metric further underscores the model's capability to provide reliable estimations for assessing drug safety. The promising performance of the multivariate interpolation model highlights its potential as a valuable computational tool in drug toxicity assessment and decision-making processes during drug development.

Diversity and Distribution Analysis in Chemical Space

Diversity and distribution analysis in chemical space refers to characterizing the variety and arrangement of molecules within a multidimensional space defined by their structural and chemical properties. It involves assessing the coverage and dispersion of molecules in this space, aiming to understand their relationships, similarities, and differences.

Diversity analysis focuses on measuring and quantifying the variety of molecules in a dataset. The analysis aims to clarify the distribution of a broad range of structural features, functional groups, and physicochemical properties of molecules in a dataset. Different diversity metrics can be used to assess the dissimilarity or similarity between molecules. We used the Tanimoto similarity index for the diversity analysis. The Tanimoto similarity is calculated based on the presence or absence of specific structural features or molecular descriptors in two compounds [25]. By evaluating diversity, we can determine if the dataset adequately represents the chemical space of interest. Upon calculating the similarity value, we obtained a Tanimoto coefficient of 0.174. This result indicates a substantial chemical diversity, as it is closer to 0 within the range of Tanimoto similarity values.

Distribution analysis involves examining the arrangement and clustering of molecules within chemical space. It aims to identify regions that are densely populated with molecules, as well as sparse or unexplored regions. Distribution analysis can be performed using visualization techniques. These methods help visualize the distribution patterns and identify clusters or subgroups of molecules with similar characteristics. We used chemical space mapping for chemical space distribution via molecular MW and AlogP values of each compound. We can explore the relationships between variables and patterns for MW and AlogP values in Figure 5.

The examination of MW reveals that the lowest observed value is 46.04, while the highest value recorded is 1201.84. This wide range of MW values indicates the presence of diverse molecular sizes within the dataset. Additionally, the analysis of AlogP demonstrates a range spanning from -9.3091 to 4.1574. The observed variation in AlogP values signifies a wide diversity of hydrophobicity or

lipophilicity among the molecules. Interestingly, the ranges for MW and AlogP exhibit similar patterns, suggesting that these properties are correlated and within the same chemical domain. The similarity in their ranges indicates that molecules with different MWs also possess a diverse range of AlogP values, implying that their hydrophobic or lipophilic characteristics are not dependent solely on their MW. This finding has significant implications in various scientific domains, particularly in drug discovery. Understanding the relationship between MW and AlogP allowed us to assess the chemical space more comprehensively, enabling the design and selection of compounds with desired molecular properties. Furthermore, this knowledge aids in exploring structure-activity relationships and identifying molecular scaffolds or substructures that contribute to specific MW and AlogP ranges.

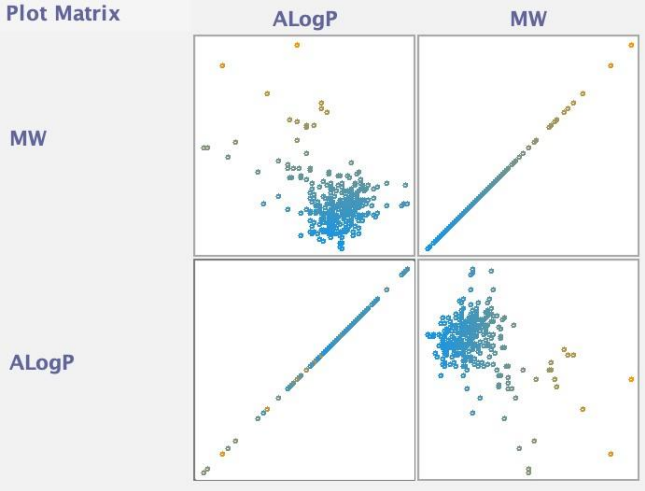

Figure 5. Plot matrix for MW and AlogP

The Comprehensive Data Regarding the Descriptors Chosen for Our Optimal Model

The process of selecting attributes is a crucial stage in machine learning modeling, as it involves identifying the significant descriptors of chemicals to attain optimal performance. Prediction models are created by utilizing various combinations of features in the descriptor pool [26]. We constructed the best-performing model using the identifiers with the highest success rate. In this study, we conducted trials with different combinations of descriptors and developed our most robust model by utilizing two specific descriptors that exhibited the highest predictive capacity. Our top model for predicting the acute oral LD_{50} value range of pharmaceuticals in mice incorporated the identifiers from the Constitutional Descriptors and Molecular Properties classes.

Constitutional Descriptors are commonly employed in QSTR modeling studies [27]. In our mathematical model, the MW is one of the most important attributes among the Constitutional Descriptors. MW represents the mass of a molecule and provides information about its size and structural complexity. The MW of a compound plays a significant role in determining its pharmacokinetic/toxicokinetic properties. As a result, the MW serves as an important factor in comprehending the behavior of a molecule within the biological system [19].

Molecular Properties identifiers have been employed in QSTR modeling studies specifically targeting acute oral toxicity in rodents [28]. Another significant descriptor in our model, the ALogP descriptor, belongs to the Molecular Properties class. ALogP stands for the predicted logarithm of the partition coefficient between octanol and water. The partition coefficient measurement determines a compound's distribution between the hydrophobic and hydrophilic phases. AlogP provides information on the potential of a compound to accumulate in adipose tissue and its permeability across biological barriers [5]. Several studies in the field of QSTR have established a correlation between chemical properties related to solubility in water or lipids and the toxicological effects of compounds [29]. In line with this, descriptive data associated with lipophilicity have been demonstrated to contribute to the

prediction models of acute toxicity. Based on a comprehensive literature analysis and our findings, we claimed that MW and AlogP are fundamental factors significantly contributing to establishing and refining acute oral toxicity models. This relationship could be attributed to processes such as absorption, excretion, and bioaccumulation of chemicals in tissues.

The Strengths and Limitations of the Optimal Model

In conventional acute toxicity studies, various compound doses are administered to experimental subjects before determining the LD_{50}/LC_{50} value of a new drug molecule. Due to the lack of knowledge regarding the toxicological effects of the novel molecule, a broad spectrum of doses can be employed. In this process, animal experiments are performed for each dose until an optimal dose is determined. The Organization for Economic Co-operation and Development (OECD) has established three acute oral toxicity procedures that rely on the utilization of experimental animals. These procedures, known as OECD-420 Fixed Dose Procedure, OECD-423 Acute Toxic Class Method, and OECD-425 Up-and-Down-Procedure, serve as standardized approaches for assessing the acute oral toxicity of substances. The reason for the publication of multiple procedures is to reduce the number of animals used and also to provide a more accurate prediction of acute toxicity. Today, ongoing research aims to minimize the use of animal models in acute toxicity testing. In this context, the OECD has published the "Acute Oral Toxicity: OECD-425 Up-and-Down Procedure" to substitute conventional acute toxicity tests with approaches that involve a reduced number of laboratory animals [30]. Adopting the perspectives of health authorities regarding the reduction of animal experimentation, we aimed to conduct preliminary studies of acute toxicity testing using mathematical models. Before conducting animal experiments to determine the LD_{50}/LC_{50} range, we propose the implementation of preliminary mathematical trials similar to our model for dose adjustment. By employing mathematical methods, dose reduction can be achieved to ensure drug safety, and ultimately, a final LD_{50}/LC_{50} value can be established through animal experiments. As a result, the use of laboratory animals can be significantly reduced while ensuring drug safety.

From a technical point of view, enhanced prediction accuracy is one of the advantages of using the interpolation technique for acute drug toxicity, specifically predicting drug LD_{50}/LC_{50} with computational modeling. Researchers can refine and optimize the model using multivariate interpolation, improving prediction accuracy for drug LD_{50}/LC_{50} values. This enables more precise assessments of a drug's acute toxicity potential, aiding in early-stage drug development and regulatory decision-making. Another advantage we can count on is cost and time efficiency. Computational modeling offers a more time and cost-effective alternative to traditional experimental methods for determining drug LD_{50}/LC_{50}. Our mathematical approach allows researchers to streamline the modeling process, reducing the need for extensive and expensive animal testing, saving resources, and accelerating drug evaluation timelines. One of the most important advantages of using this technique is reduced reliance on animal testing. This technique contributes to the reduction of animal testing in toxicological research. Using computational modeling, researchers can minimize the ethical concerns associated with animal experimentation, promoting more humane research practices while maintaining scientific rigor.

There are also disadvantages besides the advantages of using an interpolation technique for acute toxicity. The greatest challenge in applying mathematical modeling is the complexity and expertise requirements. Multivariate interpolation for drug LD_{50}/LC_{50} prediction involves complex mathematical modeling techniques. It requires expertise in computational modeling and statistical analysis, which may limit accessibility for researchers without the necessary skills or resources. Continuous model improvement is another point to note. Using a mathematical equation necessitates ongoing efforts to improve and validate the model. This includes incorporating new data, refining the model's parameters, and accounting for evolving scientific knowledge. Sustaining a robust and up-to-date model requires continuous research and resource allocation.

Our model is specifically designed to evaluate acute toxicity through oral administration in mice. However, since LD_{50} values can vary for the same molecule across different exposure routes, such as dermal or inhalation [13], the applicability of our model is limited in those situations. Furthermore, considering the species-specific toxicity variations, separate model scenarios should be developed for guinea pigs, rabbits, rats, or other experimental animal species.

Limiting our study to drug molecules presents both advantages and disadvantages. The presence of a well-balanced dataset, encompassing compounds with diverse physicochemical properties, reduces the occurrence of molecules outside the AD, consequently enhancing the model's predictive performance. By exclusively focusing on pharmaceuticals, we ensured a dataset with homogeneity. Nevertheless, our model overlooked the evaluation of non-pharmaceutical substances. QSTR models discussed in the literature regarding acute toxicity encompass a broad range of chemicals [31]. In contrast, our model is specifically designed for pharmaceutical molecules, which sets it apart in scope and focus.

The primary objective of modeling studies is to construct a dataset encompassing a wide range of molecules, maximizing its inclusiveness [4]. While we acknowledge the validity of this approach, we argue that it is equally important for molecules to belong to specific chemical groups to establish a reliable prediction model. The selection of descriptors based on specific chemical groups can pave the way for future molecule development studies. Considering that there are studies in the literature evaluating various chemicals for determining LD_{50}/LC_{50} values, our study, which solely focuses on the LD_{50}/LC_{50} values of drugs, takes an innovative approach.

Acute toxicity effects are complex processes arising from various biokinetic, cellular, and molecular events. Attempting to condense the intricate physiological phenomena associated with acute toxicity into a single numerical value may result in the loss of valuable information. Moreover, available data on LD_{50}/LC_{50} values exhibit significant variability due to variations in experimental protocols, animal species, strains, and laboratories. This variability undermines the reliability and reproducibility of acute toxicity measurements. Consequently, these challenges complicate the modeling process and lead to a relatively limited number of QSTR models for predicting acute oral toxicity compared to other endpoints [8]. However, the disadvantage mentioned in this section applies not only to mathematical modeling studies but also to animal experiments, where the LD_{50}/LC_{50} value is traditionally determined. LD_{50}/LC_{50} values have been used to initially assess relative toxicity among chemicals [4]. This issue can be addressed by integrating non-animal-based prediction models and diverse animal models and incorporating various exposure scenarios. It is worth noting that the dataset we used for our study lacks inorganic chemicals and salt structures, which could be an area for improvement in future research. As a result, our models could not provide predictions for these substances. The substances currently utilized as active drug ingredients were excluded from the evaluation.

In conclusion, the LD_{50}/LC_{50} value, representing the dosage at which 50% of specific test subjects experience fatality, is critical for assessing acute toxicity during drug development. The LD_{50}/LC_{50} test assesses the toxic effects of drugs on human health, establishing appropriate dosage regimens and ensuring their safe usage. Due to ethical considerations, traditional animal-based methods in acute toxicology studies are being replaced by mathematically based approaches. Our model has successfully predicted the five toxicologic endpoints of regulatory significance related to the acute oral toxicity of pharmaceuticals in mice. The endpoints are critical to regulatory regimes since it serves as the foundation for chemical toxicological categorization. We have argued that the current mathematical approach holds promise in assessing the LD_{50}/LC_{50} value of drug candidates during the early stages of drug development. This means new pharmaceuticals can be synthesized more cost-effective, timely, and safely. Cutting-edge models, such as ours, have the remarkable potential to significantly reduce the necessity for animal testing in toxicological research, thereby addressing ethical concerns. Reliable and validated *in silico* techniques can be utilized as an initial step in calculating the LD_{50}/LC_{50} range of drugs, serving as a valuable tool in early toxicity assessment. In conclusion, the presented mathematical model offers a reliable and practical means for estimating the LD_{50}/LC_{50} values of drugs in mice.

AUTHOR CONTRIBUTIONS

Concept: G.K., F.K.Ç.; Design: G.K., F.K.Ç.; Control: G.K., F.K.Ç.; Sources: G.K., F.K.Ç.; Materials: G.K., F.K.Ç.; Data Collection and/or Processing: F.K.Ç.; Analysis and/or Interpretation: G.K., F.K.Ç.; Literature Review: G.K., F.K.Ç.; Manuscript Writing: G.K., F.K.Ç.; Critical Review: G.K., F.K.Ç.; Other: F.K.Ç.

CONFLICT OF INTEREST

The authors declare that there is no real, potential, or perceived conflicts of interest for this article.

ETHICS COMMITTEE APPROVAL

The authors declare that the ethics committee approval is not required for this study.

REFERENCES

1. United Nations Web site. (2007). Globally Harmonized System of Classification and Labelling of Chemicals (GHS), ST/SG/AC.10/30/Rev.2. https://unece.org/. Access date: 03.05.2023.
2. Food and Drug Administration (FDA).
3. Akkaya, H., Kelleci Çelik, F. (2021). Hayvan Deneylerine Etik Açıdan Bakış. Atatürk Üniversitesi Yayınları, Erzurum, p. 75.
4. Gadaleta, D., Vuković, K., Toma, C., Lavado, G.J., Karmaus, A.L., Mansouri, K., Kleinstreuer, N.C., Benfenati, E., Roncaglioni, A. (2019). SAR and QSAR modeling of a large collection of LD_{50} rat acute oral toxicity data. Journal of Cheminformatics, 11(1), 58. [CrossRef]
5. Karaduman, G., Kelleci Çelik, F. (2023). 2D-Quantitative structure-activity relationship modeling for risk assessment of pharmacotherapy applied during pregnancy. Journal of Applied Toxicology, 43(10), 1436-1446. [CrossRef]
6. Rasulev, B., Kusić, H., Leszczynska, D., Leszczynski, J., Koprivanac, N. (2010). QSAR modeling of acute toxicity on mammals caused by aromatic compounds: The case study using oral LD50 for rats. Journal of Environmental Monitoring, 12(5), 1037-1044. [CrossRef]
7. Ruiz, P., Begluitti, G., Tincher, T., Wheeler, J., Mumtaz, M. (2012). Prediction of acute mammalian toxicity using QSAR methods: a case study of sulfur mustard and its breakdown products. Molecules, 17(8), 8982-9001. [CrossRef]
8. Lapenna, S., Gatnik, M.F., Worth, A.P. (2010). Review of QSAR models and software tools for predicting acute and chronic systemic toxicity. Publications Office of the European Union, Luxembourg, JRC61930, 1-35.
9. Schaper, M.M., Thompson, R.D., Weil, C.S. (1994). Computer programs for calculation of median effective dose (LD50 or ED50) using the method of moving average interpolation. Archives Toxicology, 68, 332-337. [CrossRef]
10. Lin, Z., Chou, W.C. (2022). Machine learning and artificial intelligence in toxicological sciences. Toxicological Sciences, 189(1), 7-19. [CrossRef]
11. Lane, T.R., Harris, J., Urbina, F., Ekins, S. (2023). Comparing LD_{50}/LC_{50} machine learning models for multiple species. ACS Chemical Health and Safety, 30(2), 83-97. [CrossRef]
12. Burden, R.L., Faires, J.D. (1997). Numerical analysis (6th ed.). Brooks/Cole Pub.
13. The European Chemicals Agency (ECHA). Web site. From https://echa.europa.eu/. Access date: 03.05.2023.
14. European Food Safety Authority (EFSA). Web site. From https://www.efsa.europa.eu/en. Access date: 03.05.2023.
15. National Library of Medicine (NLM). Web site. From https://www.nlm.nih.gov/. Access date: 03.05.2023.
16. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E. (2023). PubChem 2023 update. Nucleic Acids Research, 51(D1), D1373-D1380. [CrossRef]
17. The United States Environmental Protection Agency (U.S. EPA) Web site. (2020). User's guide for T. E. S. T. (Toxicity Estimation Software Tool) version 5.1 a java application to estimate toxicities and physical properties from molecular structure. From https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test. Accessed date: 03.05.2023.
18. Miličević, A., Šinko, G. (2022). Evaluation of the key structural features of various butyrylcholinesterase inhibitors using simple molecular descriptors. Molecules, 27(20), 6894. [CrossRef]
19. OECD (2010), Test No. 417: Toxicokinetics, OECD guidelines for the testing of chemicals, Section 4, OECD Publishing, Paris. [CrossRef]
20. Akturk, S.O., Tugcu, G., Sipahi, H. (2022). Development of a QSAR model to predict comedogenic potential of some cosmetic ingredients. Computational Toxicology, 21, 100207. [CrossRef]
21. Bojanov, B., Xu, Y. (2003). On polynomial interpolation of two variables. Journal of Approximation Theory, 120(2), 267-282. [CrossRef]

22. Hust, J.G., McCarty, R.D. (1967). Curve-fitting techniques and applications to thermodynamics, Cryogenics, 7(1), 200-206. [CrossRef]
23. Mehari, Y. (2017). Easy way to find multivariate interpolation. International Journal of Emerging Trends in Science and Technology, 4(5), 5189-5193.
24. Karaduman, G., Yang, M. (2022). An alternative method for SPP with full rank (2,1)-block matrix and nonzero right-hand side vector. Turkish Journal of Mathematics, 46(4), 1330-1341. [CrossRef]
25. OECD (2017). Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)Sar] Models. In: OECD Series on Testing and Assessment. OECD Publishing, Paris, 1-154. [CrossRef]
26. Demisse, G.B., Tadesse, T., Bayissa, Y. (2017). Data mining attribute selection approach for drought modeling: A case study for Greater Horn of Africa. International Journal of Data Mining and Knowledge Management Process, 7(4), 1-16. [CrossRef]
27. Kelleci Çelik, F., Karaduman, G. (2022). In silico QSAR modeling to predict the safe use of antibiotics during pregnancy. Drug and Chemical Toxicology, 46(3), 1-10. [CrossRef]
28. Devillers, J. (2004). Prediction of mammalian toxicity of organophosphorus pesticides from QSTR modeling. SAR and QSAR in Environmental Research, 15(5-6), 501-510. [CrossRef]
29. Abraham, M.H., Grellier, P.L., Kamlet, M.J., Doherty, R.M., Taft, R.W., Abboud, J.L.M. (1989). The use of scales of hydrogen-bond acidity and basicity in organic chemistry. Revista Portuguesa de Química, 31, 85.
30. OECD (2022), Test No. 425: Acute Oral Toxicity: Up-and-Down Procedure, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris. [CrossRef]
31. Zhu, H., Martin, T.M., Ye, L., Sedykh, A., Young, D.M., Tropsha, A. (2009). Quantitative structure activity relationship modeling of rat acute toxicity by oral exposure. Chemical Research in Toxicology, 22(12), 1913-1921. [CrossRef]