HIGH SCHOOL EDITION
Journal of Student Research

Development of Hybrid Machine Learning Model to Predict Intrinsic Clearance for Pesticides

Kyuri Lim

Bergen County Technical High School, USA

ABSTRACT

Predicting intrinsic clearance (Clint) is essential for understanding the pharmacokinetics of pesticides, as it directly influences dosing regimens and the overall behavior of chemicals within biological systems. This study aimed to develop and validate a hybrid machine learning model to accurately predict Clint for various pesticide categories, utilizing publicly available data from the U.S. Environmental Protection Agency's National Center for Computational Toxicology (EPA NCCT) High-Throughput Toxicokinetics (HTTK) dataset. Molecular descriptors were calculated using the PaDEL software, and relevant descriptors were selected using the k-nearest neighbors (kNN) algorithm based on their correlation with Clint values. Three machine learning models—Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN)—were trained and evaluated across four pesticide categories: Total, Herbicides, Insecticides, and Fungicides. The Random Forest model achieved the highest R^2 value of 0.967 for Herbicides, while XGBoost outperformed the other models for Insecticides (R^2=0.806) and Fungicides (R^2=0.840), as well as the Total category (R^2=0.744). Despite these promising results, the study faced limitations such as the treatment of outliers, the presence of excessive zeros in the dataset, and small sample sizes (e.g., n=20 for Herbicides), which could impact the accuracy of the models. The hybrid model, by selecting the optimal algorithm based on input chemical structures, demonstrates significant potential in predicting Clint for new chemicals, offering a rapid and reliable alternative to traditional in vivo methods. These findings contribute to the field of computational toxicology by enhancing the predictive capabilities of in silico models and supporting the development of safer chemical compounds.

Introduction

Intrinsic clearance (Clint) is a critical pharmacokinetic parameter that reflects the liver's ability to metabolize drugs independent of blood flow. Accurate estimation of Clint is essential for predicting a drug's clearance from the body, determining appropriate dosing regimens, and understanding the pharmacokinetic behavior of compounds (Obach, 1999). Traditional approaches to estimating Clint have relied heavily on in vivo studies using animal models. While these studies provide valuable data, they are costly, time-consuming, and raise ethical concerns (Brown et al., 2007). Additionally, the differences in drug metabolism between animals and humans limit the applicability of these methods, necessitating the development of alternative approaches (Zhang et al., 2010).

Recent advancements in computational methods have led to the emergence of in silico models, which aim to reduce or replace animal testing by utilizing computational tools, including machine learning (ML) algorithms, to predict intrinsic clearance based on molecular descriptors and other biochemical properties (van de Waterbeemd & Gifford, 2003). In silico models offer significant advantages, including faster predictions, reduced costs, and the elimination of ethical concerns associated with animal testing (Patilea-Vrana & Unadkat, 2018). Moreover, these models can provide insights into drug metabolism that are more directly relevant to human physiology (Alqahtani et al., 2013). However, developing reliable in silico models for Clint prediction poses several challenges. A major issue is the selection of relevant molecular descriptors from a vast pool of potential features. The performance of these models depends heavily on the quality and relevance of the input data, which must be carefully curated to avoid issues such as outliers, non-parametric distributions, and excessive zero values (Gleeson et al., 2011). Machine learning techniques, with their ability to manage complex datasets and identify patterns, are particularly well-suited for addressing these challenges (Patilea-Vrana & Unadkat, 2018). Given the diverse strengths of different ML algorithms, it is important to evaluate multiple techniques to identify the most effective method for Clint prediction. This study focused on three prominent ML methods: Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN). By comparing the performance of these models across different categories of pesticides, Total, Herbicides, Insecticides, and Fungicides, this research aimed to determine the best-performing algorithm for each category. Furthermore, the concept of a hybrid model is explored, which integrates the predictions from the top-performing models to enhance overall predictive accuracy (Jones et al., 2013).

The primary objective of this study was to develop and validate in silico models for predicting intrinsic clearance using a combination of ML techniques. By systematically comparing RF, XGBoost, and ANN, and by exploring the potential of a hybrid model, this research aimed to contribute to the field of pharmacokinetics by providing a reliable, efficient, and ethical alternative to traditional methods. The outcomes of this study are expected to advance the development of safer and more effective pharmaceutical compounds, offering significant benefits to both scientific research and practical drug development (Obach, 1999).

Methodology

Training Data: EPA Rat Intrinsic Clearance *In Vitro* Study Data

The study utilized intrinsic clearance (Clint) data from the U.S. Environmental Protection Agency (EPA) National Center for Computational Toxicology's High-Throughput Toxicokinetics (HTTK) training dataset. This dataset provides a comprehensive analysis of the intrinsic clearance of various compounds in rat models, offering valuable insights into their metabolic characteristics. The training dataset includes detailed information such as chemical classifications, molecular formulas, and key pharmacokinetic parameters, including intrinsic clearance. This dataset serves as the foundation for developing computational models to predict how chemicals are metabolized in vivo.

Calculation of Molecular Descriptors Using PaDEL Modeling and kNN

To generate the molecular descriptors needed for model development, the PaDEL software was used. PaDEL is an open-source tool capable of calculating a wide range of molecular descriptors and fingerprints from chemical structures. These descriptors quantify numerous structural properties of molecules, such as bond interactions, atom counts, and molecular topology, providing essential inputs for the machine learning models. The primary input required for PaDEL is the chemical structure, typically provided in the form of a SMILES (Simplified Molecular Input Line Entry System) code or other chemical structure formats like SDF (Structure-Data File) or MOL files. These inputs allow PaDEL to compute various descriptors, including bond interactions, atom counts, and molecular topology, which are essential for feeding into the machine learning models for predicting intrinsic clearance. Once the molecular descriptors were calculated, they were combined with the intrinsic clearance (Clint) data, where the Clint values served as the dependent variable (y), and the calculated descriptors formed the independent variable group (X). This combined dataset was then used as the input for the machine learning models to predict intrinsic clearance.

Following this, the k-nearest neighbors (kNN) algorithm was employed to further refine the dataset. The kNN algorithm classifies data points based on the majority class of their nearest neighbors in the descriptor space. Additionally, kNN ranks these descriptors by examining their influence on the classification of neighbors, allowing for the identification of the most relevant descriptors. This process significantly narrowed down the list from the original 1,444 molecular descriptors, focusing on those most likely to contribute to accurate predictions.

The kNN algorithm operates based on the principle of similarity, where the class of a given data point is determined by the majority class among its *k* nearest neighbors. The classification decision is made according to the equation:

$$f(x) = \frac{1}{k}\sum_{i\in N,(x)} (\mathrm{y_i})$$

Where:

- $f(x)$ is the predicted output for the input xxx,
- $N_k(x)$ denotes the set of the k nearest neighbors of x,
- y_i is the observed class or value of the neighbor.

Machine Learning Techniques: Random Forest, XGBoost, and ANN

In this study, three machine learning models were employed: Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN). Each model was chosen for its unique strengths and applicability to different aspects of the dataset:

Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions to produce a final output. The algorithm works by selecting a random subset of features for each tree, thus creating a diverse set of trees and reducing overfitting. The final prediction is made by averaging the predictions of all trees (in the case of regression) or by majority voting (in the case of classification).

The prediction of a Random Forest model for a regression problem is given by:

$$y\hat{} = \frac{1}{N}\sum_{n=1}^{N} T_n(x)$$

Where:

- y^ is the predicted value,
- N is the number of trees in the forest,
- $T_n(x)$ is the prediction of the n-th tree for input x.

XGBoost

XGBoost is a powerful gradient boosting algorithm that builds models sequentially, with each new model correcting the errors of its predecessor. This iterative process leads to highly accurate and efficient predictions. XGBoost optimizes a specific objective function and applies regularization to reduce overfitting, making it a robust choice for both regression and classification tasks.

The prediction for an input xxx in XGBoost is given by:

$$y\hat{} = \sum_{k=1}^{k} f_k(x),\ f_k\ \in \rho$$

Where:

- y^ is the predicted value,
- k is the number of trees,

- $f_k(x)$ is the prediction of the k-th tree,
- ρ denotes the space of regression trees.

The objective function L optimized during training includes both a loss function and a regularization term:

$$L(\emptyset) = \sum_{i} l(y_k\, , y^\wedge{}_k\,) + \sum_{k=1} \Omega(f_k)$$

Where:

- l is the loss function (e.g., mean squared error),
- $\Omega(f_k)$ is the regularization term for tree k.

Artificial Neural Networks (ANN)

ANN is a deep learning approach that simulates the neural networks of the human brain, passing data through multiple layers of interconnected nodes (neurons). Each layer captures different aspects of the data, allowing the network to learn complex patterns. The model adjusts its internal connections (weights and biases) during training to minimize the error between predicted and observed values.

The output of a neuron in a neural network is computed as:

$$a_j^{(l)} = \sigma \sum_{i=1}^{n^{(l-1)}} w_{ji}^{(l)} a_j^{(l-1)} + b_j^{(l)}$$

Where:

- $a_j^{(l)}$ is the activation of the j-th neuron in layer l,
- σ is the activation function (e.g., ReLU, sigmoid),
- $w_{ji}^{(l)}$ is the weight connecting neuron i in layer l-1 to neuron j in layer l,
- $b_j^{(l)}$ is the bias term for neuron j in layer l.

Model Validation and Hybrid Machine Learning

The performance of each model was evaluated using the coefficient of determination (R^2) from simple multiple linear regression between observed and predicted intrinsic clearance values. The R^2 value serves as a key indicator of how well the model's predictions match the observed data, with higher R^2 values indicating better model performance. This approach provided a straightforward and effective means of assessing the accuracy of the models across different pesticide categories.

To optimize predictive accuracy, a hybrid model approach was explored. The hybrid model was constructed by selecting the best-performing model for each pesticide category based on its R^2 value. The predictions from these models were then integrated to create a combined output that leveraged the strengths of each individual model. This approach was particularly effective in enhancing overall predictive accuracy by ensuring that the most suitable model was applied to each specific category.

This hybrid model approach is especially advantageous for predicting intrinsic clearance (Clint) for newly discovered chemicals. Once the molecular descriptors are generated from the chemical structure, provided as a SMILES (Simplified Molecular Input Line Entry System) code, the hybrid model can dynamically select and apply the most accurate algorithm for the corresponding category. This ensures that Clint predictions are as reliable and precise as possible, making this approach a powerful tool for rapid evaluation in chemical discovery and safety assessments.

Results

Analysis of Training Data and Parameter Selection Using kNN

Figure 1 illustrates the top 10 molecular descriptors for each pesticide category—Total, Herbicides, Insecticides, and Fungicides—ranked by their correlation with intrinsic clearance (Clint) values, as determined by the k-nearest neighbors (kNN) algorithm. These rankings showcase the descriptors most strongly associated with Clint within each category. While Figure 1 highlights the top 10 descriptors based on their ranking, it's important to note that in the subsequent machine learning models, only descriptors with an R^2 value greater than 0.5 were selected as denoted in Table 1. This careful selection process ensures that the models focus on the most predictive molecular features, thereby enhancing the accuracy of Clint predictions (Obach, 1999; Saeed et al., 2017).

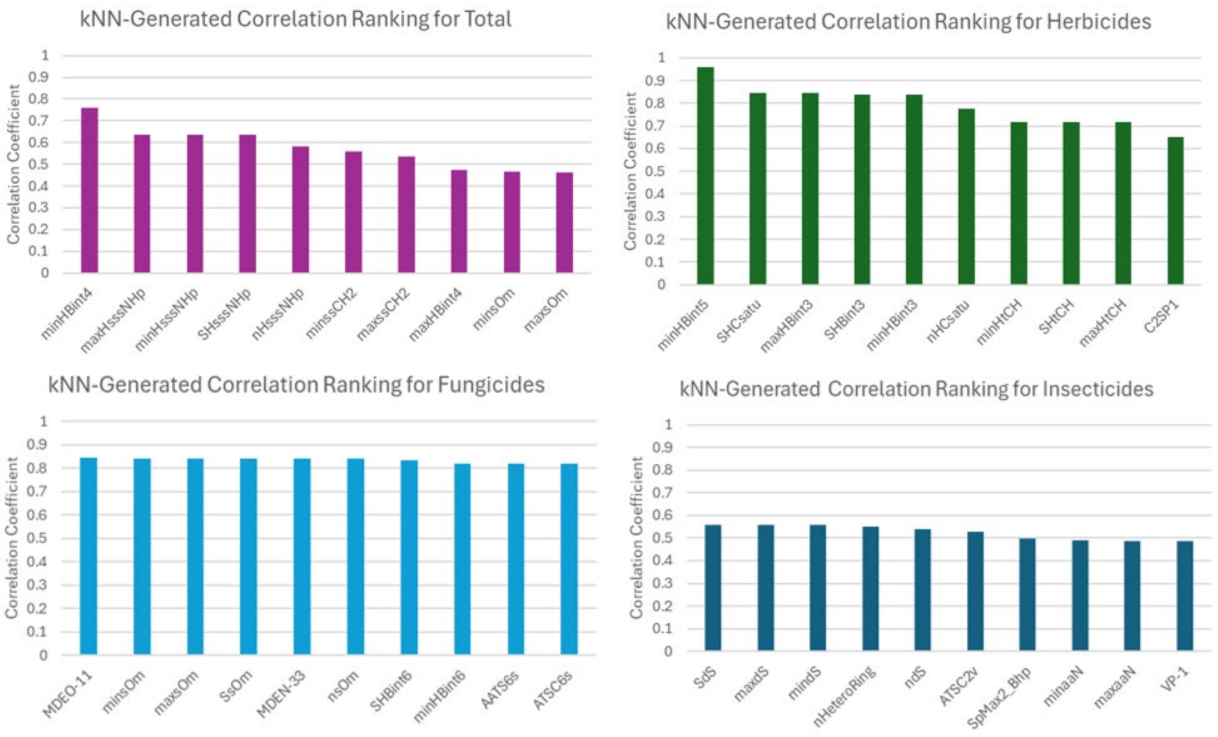

Figure 1. kNN-Generated Correlation Ranking for Total (top left), Herbicides (top right), Fungicides (bottom left) and Insecticides (bottom right).

Table 1. Selected Descriptors (R^2>0.5) for Each Pesticide Category Based on kNN Analysis

Pesticide Category	**Selected Descriptors**	**R^2 Value Range**
Total	minHBint4, maxHsssNHp, minHsssNHp, SHsssNHp, nHsssNHp	0.580 - 0.758
Herbicides	minHBint5, SHCsatu, maxHBint3, SHBint3, minHBint3, nHCsatu, minHtCH, SHtCH, maxHtCH, C2SP1	0.837 - 0.960
Insecticides	SdS, maxdS, mindS, nHeteroRing, ndS	0.538 - 0.557

Fungicides	MDEO-11, minsOm, maxsOm, SsOm, MDEN-33	0.838 - 0.845

Explanation of High Correlations for Each Pesticide Group

The selected molecular descriptors exhibited strong correlations with intrinsic clearance (Clint) values across the different pesticide categories—Total, Herbicides, Insecticides, and Fungicides. These correlations can be explained by the specific roles that these descriptors play in the metabolic processes affecting Clint. Below is a detailed analysis of why these descriptors are significant for each pesticide group. Table 2 provides a detailed overview of the molecular descriptors that were found to have the highest correlations with intrinsic clearance (Clint) across the four pesticide categories. The table includes both the descriptor names and their respective definitions, offering insights into why these specific molecular features are significant in predicting Clint for each group.

Table 2. Description of molecular descriptors in Top 10 Ranks.

Total		**Herbicides**	
minHBint4	The minimum value of hydrogen bond interactions involving four atoms.	minHBint5	The minimum value of hydrogen bond interactions involving five atoms.
maxHsssNHp	The maximum value of three single-bonded hydrogens attached to a nitrogen with a lone pair.	SHCsatu	The sum of the contributions from all carbon atoms with single bonds only in the molecule
minHsssNHp	The minimum value of the three single-bonded hydrogens attached to a nitrogen with a lone pair.	maxHBint3	The maximum value of hydrogen bond interactions involving three atoms.
SHsssNHp	The sum of the values for three single-bonded hydrogens attached to a nitrogen with a lone pair.	SHBint3	The sum of all hydrogen bond interactions involving three atoms.
nHsssNHp	The count of occurrences of three single-bonded hydrogens attached to a nitrogen with a lone pair.	minHBint3	The minimum value of hydrogen bond interactions involving three atoms.
minssCH2	The minimum value of a carbon atom attached to two hydrogens and two single bonds.	nHCsatu	The count of carbon atoms with only single bonds in the molecule.
maxssCH2	The maximum value of a carbon atom attached to two hydrogens and two single bonds.	minHtCH	The minimum value of hydrogen atoms attached to a carbon atoms with a triple bond.
maxHBint4	The maximum value of hydrogen bond interactions involving four atoms.	SHtCH	The sum of contributions from all hydrogen atoms attached to carbons with triple bonds.
minsOm	The minimum value of oxygen atoms in a specific molecular environment.	maxHtCH	The maximum value of hydrogen atoms attached to a carbon atoms with a triple bond.
maxsOm	The maximum value of oxygen atoms in a specific molecular environment.	C2SP1	The count of carbon atoms with two attached single bonds and one attached triple bond.
Insecticides		**Fungicides**	
SdS	The sum of all sulfur-sulfur single bonds in a molecule.	MDEO-11	The mean distance between oxygen atoms separated by 11 bonds within the molecule.
maxdS	The maximum value of any sulfur atom connected by a single bond in a molecule.	minsOm	The minimum value of oxygen atoms in a specific molecular environment.
mindS	The minimum value of any sulfur atom connected by a single bond in a molecule.	maxsOm	The maximum value of oxygen atoms in a specific molecular environment.

nHeteroRing	The number of non-carbon atoms present in a ring structure within the molecule.	SsOm	The sum of contributions from oxygen atoms in a specific molecular environment.
ndS	The count of sulfur atoms connected by a single bond in a molecule.	MDEN-33	The mean distance between nitrogen atoms separated by 33 bonds within the molecule.
ATSC2V	The autocorrelation of the topological structure at lag 2 using van der Waals volumes as weights.	nsOm	The count of oxygen atoms in a specific molecular environment.
SpMax2_Bhp	The maximum value of the Burden eigenvalue for a 2-bond length path in the molecule, focusing on polar hydrogens.	SHBint6	The sum of all hydrogen bond interactions involving six atoms.
minaaN	The minimum value of NH2 or NH in the molecule.	minHBint6	The minimum value of hydrogen bond interactions involving six atoms.
maxaaN	The maximum value of NH2 or NH in the molecule.	AATS6s	The average autocorrelation of the topological structure at lag 6 using atom-level van der Waals surface areas as weights.
VP-1	The first component of the VAMP descriptor, related to the molecular volume and shape.	ATSC6s	The centered autocorrelation of the topological structure at lag 6 using atom-level van der Waals surface areas as weights.

Total

For the Total category, the descriptors minHBint4, maxHsssNHp, minHsssNHp, SHsssNHp, and nHsssNHp are related to hydrogen bonding, particularly involving nitrogen atoms with lone pairs. Hydrogen bonding is crucial for the interaction between compounds and metabolic enzymes, such as cytochrome P450, which significantly influences a molecule's metabolism and clearance rate (Lewis, 2002; Obach, 1999). The descriptors minssCH2 and maxssCH2 reflect specific carbon environments that may impact the compound's reactivity and interaction with these enzymes (Smith et al., 1996). Additionally, minsOm and maxsOm quantify the presence and environment of oxygen atoms, which are critical for oxidation reactions during metabolism, further affecting Clint (van de Waterbeemd & Gifford, 2003).

Herbicides

In the Herbicides category, descriptors such as minHBint5, maxHBint3, SHBint3, and minHBint3 focus on hydrogen bond interactions involving three or five atoms. The strength and number of these hydrogen bonds directly affect how herbicides are metabolized, thus influencing their clearance rates (Obach, 1999; Gleeson, 2008). The descriptors SHCsatu and nHCsatu are related to the saturation of carbon atoms, which affects metabolic stability and oxidation processes in herbicides (Hansch et al., 2005). Descriptors like minHtCH, maxHtCH, and SHtCH involve hydrogen atoms attached to carbons with triple bonds, which can impact the molecule's rigidity and susceptibility to metabolism, thereby altering its clearance (Smith et al., 1996). C2SP1 represents specific carbon bonding environments that influence the molecule's overall shape and electron distribution, critical for interaction with metabolic enzymes (Klopman & Chakravarti, 2003).

Insecticides

For Insecticides, descriptors SdS, maxdS, mindS, and ndS are related to sulfur-sulfur single bonds, significant because sulfur atoms are key sites for oxidation or reduction during metabolism (Lewis, 2002; Kirchmair et al., 2012). The nHeteroRing descriptor counts non-carbon atoms in ring structures, with heteroatoms like oxygen, nitrogen, or sulfur altering the molecule's reactivity and metabolism (Gleeson, 2008). ATSC2V and SpMax2_Bhp involve topological structure and autocorrelation using van der Waals volumes, impacting how the molecule fits into the active site of enzymes (Ertl et al., 2000). Descriptors minaaN and maxaaN reflect the presence of NH2 or NH groups, common sites for metabolic reactions like deamination, which directly influence Clint (Smith et al., 1996). VP-1 relates to molecular volume and shape, factors crucial for how well a molecule is processed by enzymes, affecting its clearance (Kirchmair et al., 2012).

Fungicides

In the Fungicides category, MDEO-11 and MDEN-33 descriptors reflect the mean distance between oxygen or nitrogen atoms, which influence the molecule's 3D structure and its interaction with metabolic enzymes (Kirchmair et al., 2012). Descriptors like minsOm, maxsOm, nsOm, and SsOm focus on oxygen atoms' environments, critical for metabolic processes such as oxidation (van de Waterbeemd & Gifford, 2003). SHBint6 and minHBint6 involve hydrogen bonds with six atoms, indicating complex interactions affecting how the molecule is metabolized (Lewis, 2002). Finally, AATS6s and ATSC6s describe autocorrelation of topological structures at lag 6, analyzing the molecule's shape and electron distribution, which influence its interactions with enzymes (Ertl et al., 2000).

Machine Learning Results

Three machine learning models—Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN)—were trained and evaluated across the four pesticide categories: Total, Herbicides, Insecticides, and Fungicides. The performance of these models was assessed using the coefficient of determination (R^2), which indicates how well the model's predictions match the observed intrinsic clearance values.

Table 3 summarizes the R^2 values for each model across the different pesticide categories. The Random Forest model achieved the highest R^2 value for the Herbicides category, while the XGBoost model outperformed the others in the Total, Insecticides, and Fungicides categories.

Table 3. Correlation coefficient (R^2) to evaluate model performance of the three models.

Model	R^2 (observed vs. predicted)			
	Total (N=66)	Herbicides (N=20)	Insecticides (N=24)	Fungicides (N=22)
Random Forest	0.533	0.967	0.679	0.742
XGBoost	0.744	0.705	0.806	0.840
ANN	0.098	0.791	0.303	0.372

Hybrid Machine Learning Model

To optimize predictive accuracy, a hybrid model was developed by selecting the best-performing machine learning model for each pesticide category based on the R^2 values. As shown in Table 4, the hybrid model dynamically chooses the model that performed the best for each individual prediction task within each category. This approach ensures that the strongest predictive algorithm is applied to each new data point, thereby enhancing overall predictive accuracy across the pesticide categories. The use of hybrid models, which leverage the strengths of different algorithms to improve prediction outcomes, is well-supported in the literature (Zhang & Ma, 2012; Kourou et al., 2015).

Table 4. Multiple Linear Equation for Each Machine Learning Model with the Highest R^2.

Pesticide Category	Selected Model	R^2
Total	XGBoost	0.744
Herbicides	Random Forest	0.967
Insecticides	XGBoost	0.806
Fungicides	XGBoost	0.840

This approach can be particularly powerful when predicting intrinsic clearance (Clint) for a newly discovered chemical. After receiving the chemical structure in the form of a SMILES (Simplified Molecular Input Line Entry System) code, molecular descriptors are calculated, and the hybrid model then selects the best-performing model for the corresponding class or category to predict Clint. This streamlined process allows for rapid and accurate predictions of Clint, facilitating more efficient drug discovery and chemical safety evaluations.

Conclusion

This study successfully demonstrated the application of hybrid machine learning models to predict intrinsic clearance (Clint) for various pesticide categories, leveraging the strengths of different algorithms—Random Forest, XGBoost, and Artificial Neural Networks (ANN). By selecting the best-performing model for each pesticide category, the hybrid model approach provided a significant enhancement in predictive accuracy, particularly when compared to individual models.

The predictive accuracy varied across the different pesticide categories. For the Herbicides category, the Random Forest model achieved the highest performance with an R^2 value of 0.967, while for Insecticides and Fungicides, XGBoost was the best-performing model with R^2 values of 0.806 and 0.840, respectively. In the Total category, XGBoost also outperformed the other models, achieving an R^2 value of 0.744. The overall accuracy range for these models spanned from 0.098 (for ANN in the Total category) to 0.967 (for Random Forest in the Herbicides category).

However, the study also faced significant challenges, particularly in the treatment of outliers and the handling of datasets with an excessive number of zeros. These issues can distort model performance, leading to reduced accuracy and biased predictions. Outlier treatment required careful consideration, as outliers can disproportionately influence model training, especially in smaller datasets. Similarly, the presence of numerous zeros in the dataset, which might represent undetectable or negligible Clint values, posed difficulties in model fitting and required specific preprocessing strategies to mitigate their impact on model performance.

Despite these challenges, the hybrid model's ability to dynamically select the optimal predictive algorithm based on the input chemical structure makes it a powerful tool for rapid and accurate prediction of Clint for newly discovered chemicals. By providing a streamlined workflow—from receiving a chemical structure in the form of a SMILES code, to calculating molecular descriptors, to predicting Clint—this approach has the potential to significantly accelerate the drug discovery process and enhance chemical safety evaluations.

The study also highlighted the importance of molecular descriptors in predicting Clint, with key descriptors related to hydrogen bonding, molecular topology, and specific atom environments showing strong correlations with metabolic clearance rates across different pesticide classes. These findings underscore the value of using detailed molecular descriptors in combination with advanced machine learning techniques to improve the prediction of pharmacokinetic parameters.

The integration of in silico models, such as the one developed in this study, offers a promising alternative to traditional in vivo methods, reducing reliance on animal testing while providing more efficient and cost-effective ways to assess the metabolic clearance of chemicals. As machine learning and computational chemistry continue to evolve, further refinement and validation of these models will undoubtedly enhance their predictive power and broaden their applicability across different chemical and pharmacological domains.

Limitations

The study faced significant limitations, including the presence of outliers that could skew model predictions and reduce accuracy, particularly in smaller datasets. Additionally, the excessive number of zero values in the intrinsic clearance (Clint) measurements posed challenges for model training, potentially leading to less reliable predictions.

Moreover, the limited data availability in certain pesticide categories, such as herbicides with only 20 samples, constrained the models' ability to generalize and perform robustly, especially in models like Random Forest that typically require larger datasets to achieve optimal performance.

Acknowledgments

I would like to express my gratitude to the U.S. Environmental Protection Agency (EPA) National Center for Computational Toxicology (NCCT) for providing publicly accessible intrinsic clearance data and for making available the OPERA/PaDEL model, which was instrumental in generating the molecular descriptors used in this study.

References

Alqahtani, S., Mohamed, L. A., & Kaddoumi, A. (2013). Development of in silico models to predict intrinsic clearance of drugs. European Journal of Pharmaceutical Sciences, 48(3), 634-642. https://doi.org/10.1016/j.ejps.2012.12.007

Brown, H. S., Griffin, M., & Houston, J. B. (2007). Prediction of in vivo drug clearance from in vitro data: Experience from GlaxoSmithKline's drug metabolism and pharmacokinetics department. Current Opinion in Drug Discovery & Development, 10(5), 435-448.

Ertl, P., Rohde, B., & Selzer, P. (2000). Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. Journal of Medicinal Chemistry, 43(20), 3714-3717. https://doi.org/10.1021/jm000942e

Gleeson, M. P. (2008). Generation of a set of simple, interpretable ADMET rules of thumb. Journal of Medicinal Chemistry, 51(4), 817-834. https://doi.org/10.1021/jm701122q

Gleeson, M. P., Hersey, A., Montanari, D., & Overington, J. (2011). Probing the links between in vitro potency, ADMET, and physicochemical parameters. Nature Reviews Drug Discovery, 10(3), 197-208. https://doi.org/10.1038/nrd3367

Hansch, C., Hoekman, D., & Gao, H. (2005). Comparative QSAR: Toward a deeper understanding of chemicobiological interactions. Chemical Reviews, 105(5), 2093-2137. https://doi.org/10.1021/cr0300326

Jones, H. M., Rowland-Yeo, K., & Chien, J. Y. (2013). Application of physiologically based pharmacokinetic modeling in drug development. Clinical Pharmacology & Therapeutics, 93(5), 426-437. https://doi.org/10.1038/clpt.2013.53

Kirchmair, J., Williamson, M. J., Tyzack, J. D., Tan, L., Bond, P. J., & Glen, R. C. (2012). Computational prediction of metabolism: Sites, products, SAR, P450 enzyme dynamics, and mechanisms. Journal of Chemical Information and Modeling, 52(3), 617-648. https://doi.org/10.1021/ci200542m

Klopman, G., & Chakravarti, S. K. (2003). Screening of high production volume chemicals for estrogen receptor binding activity (II) by the MultiCASE expert system. Chemosphere, 51(6), 461-468. https://doi.org/10.1016/S0045-6535(02)00769-7

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

Lewis, D. F. V. (2002). Molecular modeling of human cytochrome P450-substrate interactions. Drug Metabolism Reviews, 34(1-2), 73-114. https://doi.org/10.1081/DMR-120001717

Obach, R. S. (1999). Prediction of human clearance of twenty-nine drugs from hepatic microsomal intrinsic clearance data: An examination of in vitro half-life approach and nonspecific binding to microsomes. Drug Metabolism and Disposition, 27(11), 1350-1359. https://doi.org/10.1124/dmd.119.086488

Patilea-Vrana, G. I., & Unadkat, J. D. (2018). Development of in silico methods to predict hepatic clearance and assess their performance. Journal of Pharmacokinetics and Pharmacodynamics, 45(1), 109-124. https://doi.org/10.1007/s10928-017-9545-0

Saeed, F., Durán, A., Korzekwa, K. R., & Nagar, S. (2017). Prediction of human clearance for drugs that are metabolized by human aldehyde oxidase. Drug Metabolism and Disposition, 45(11), 1266-1275. https://doi.org/10.1124/dmd.117.077545

Smith, D. A., Jones, B. C., & Walker, D. K. (1996). Design of drugs involving the concepts and theories of drug metabolism. Journal of Medicinal Chemistry, 39(9), 3103-3110. https://doi.org/10.1021/jm9509998

van de Waterbeemd, H., & Gifford, E. (2003). ADMET in silico modelling: Towards prediction paradise? Nature Reviews Drug Discovery, 2(3), 192-204. https://doi.org/10.1038/nrd1032

Zhang, C., & Ma, Y. (2012). Ensemble machine learning: Methods and applications. Springer Science & Business Media.

Zhang, L., Reynolds, K. S., Zhao, P., & Huang, S. M. (2010). Drug-drug interactions: Scientific and regulatory perspectives. Clinical Pharmacology & Therapeutics, 89(1), 124-128. https://doi.org/10.1038/clpt.2010.199