Artificial Intelligence Chemistry 3 (2025) 100093

Contents lists available at ScienceDirect

Artificial Intelligence Chemistry

journal homepage: www.journals.elsevier.com/artificial-intelligence-chemistry

# Machine Learning (ML)-driven quantitative structure-pharmacokinetic relationship (QSPKR) modeling of the tissue-to-plasma partition coefficient ($K_p$) of drugs across different tissues

Souvik Pore, Kunal Roy [*,1]

*Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, 188 Raja S C Mullick Road, Kolkata 700032, India*

ARTICLE INFO

*Keywords:*
Tissue-to-plasma partition coefficient (Kp)
Machine Learning (ML)
Quantitative structure-pharmacokinetic relationship (QSPKR)

ABSTRACT

In drug discovery, estimating the drug candidate's pharmacokinetic (PK) parameters is crucial for determining its safety and efficacy within the body. The tissue-to-plasma partition coefficient (Kp) indicates how a drug partitions within a tissue, potentially leading to tissue-specific activity or toxicity. Therefore, determining $K_p$ values for a drug is essential for its safety assessment. However, only a limited number of such studies are available. Here, we developed machine learning (ML)-driven quantitative structure-pharmacokinetic relationship (QSPKR) models to predict the $K_p$ values for drugs across 11 different tissues. Initially, we developed models to predict $K_p$ values for drugs with missing $K_p$ values for specific tissues within the dataset solely based on the structural and physicochemical properties of the drugs. Subsequently, another set of models was developed using both structural and physicochemical properties and the $K_p$ values from other tissues. In this case, predicted values from the initial models were also incorporated where experimental $K_p$ values were unavailable. These models demonstrate significant improvement in predictability ($Q^2_{F1} = 0.79–0.95$, $Q^2_{F2} = 0.78–0.95$) for a drug compared to the initial models. Here, we conducted a screening using a true external dataset from the SIDER database. This analysis indicates that compounds with higher tissue partitioning are more likely to exhibit toxicity to that specific tissue. Finally, a Python-based tool (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/kp-calculator) was created to predict $K_p$ values for drugs in different tissues.

## 1. Introduction

The discovery of a new drug molecule involves identifying a novel molecule with the desired properties for treating a disease. [1] Pharmacokinetic (PK) parameters are pivotal factors that influence the selection of a molecule for clinical trials. Generally, PK parameters encompass Absorption, Distribution, Metabolism, and Elimination, collectively referred to as ADME. [2] The PK of a drug is a multifaceted process that depends on multiple factors, including the physicochemical properties of the drug, blood flow rate, tissue permeability, the volume of various tissues, the affinity of tissues for a particular molecule, and tissue composition. [3] A significant number of molecules do not succeed in becoming good therapeutic agents due to their suboptimal PK properties. [4,5] Therefore, a comprehensive understanding of the PK parameters associated with a drug is essential in the drug discovery and development process, as these parameters directly affect the efficacy and safety of the drug. That is why a thorough PK profiling is essential to ensure that a drug achieves its target site at an optimal concentration for the desired duration, thereby minimizing potential toxic effects while maximizing therapeutic benefits. [6]

Characterizing and predicting PK parameters are crucial for both pharmaceutical industries and academics. Early prediction of PK parameters during drug discovery is vital, as it aids in discarding molecules with undesirable properties, thereby reducing costs and resource consumption. [7,8] In the last few years, physiologically based pharmacokinetic (PBPK) modeling has been employed to predict a drug's PK parameters. This modeling expedites the drug discovery and development process while minimizing the costs associated with necessary in vivo and in vitro experiments. [9–11] PBPK models utilize differential equations that simulate the physiological conditions a drug encounters in the body, capturing the PK processes. Multiple physiological and biochemical parameters are considered in these models to construct a

* Corresponding author.
*E-mail addresses:* kunalroy_in@yahoo.com, kunal.roy@jadavpuruniversity.in (K. Roy).
[1] ORCID https://orcid.org/0000-0003-4486-8074

https://doi.org/10.1016/j.aichem.2025.100093
Received 30 May 2025; Received in revised form 13 July 2025; Accepted 30 July 2025
Available online 31 July 2025
2949-7477/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

multi-compartment model, with each compartment representing one or more bodily organs. [11,12]

The tissue-to-plasma partition coefficient ($K_p$) is one of the important pharmacokinetic (PK) parameters, as it describes the distribution of drugs within specific tissues from systemic circulation under steady-state conditions. [13,14] It is defined as the ratio of a drug's concentration in a tissue to plasma at equilibrium. [15] Essentially, it represents the amount of drug transferred to or retained in that tissue. The extent of transfer depends on tissue partitioning and the drug's affinity for tissue composition, plasma proteins, and blood cells. [16] Because tissue composition varies throughout the body, $K_p$ values are tissue-specific. Traditionally, $K_p$ values were determined through in vivo experimentation. While in vivo studies provide accurate $K_p$ values, conducting these experiments requires a significant number of animals and resources for sample processing and analysis. [17] Determining $K_p$ involves measuring concentrations in both tissue and plasma; however, obtaining tissue concentration is considerably challenging. Additionally, routine data collection and analysis of numerous chemicals make this process costly, time-consuming, and resource-intensive. [18] Therefore, developing new methods, such as in-silico modeling, for predicting $K_p$ from more readily accessible parameters is crucial. Three main methods have been applied to predict a drug's $K_p$ value: (i) the tissue-composition-based (TCB) algorithm, developed using tissue-specific parameters and the drug's physicochemical properties, [19,20] (ii) the correlation-based algorithm, which utilizes compound-specific and in-vivo information, [21] and (iii) machine learning (ML) based non-linear algorithms. [16]

In recent years, the modeling of pharmacokinetic (PK) parameters has become a crucial alternative to experimental procedures due to the increasing computational power and resources available. Advanced modeling algorithms have been developed and implemented to predict PK parameters effectively. [22–25] Various machine learning (ML) algorithms have been employed to create models predicting PK parameters such as volume of distribution, clearance, and plasma protein binding. [26] This study focuses on developing ML-based quantitative structure-pharmacokinetic relationship (QSPKR) models to predict the Kp values of drugs across different tissues. ML, a branch of artificial intelligence (AI), is trained on previous data to enhance performance in future tasks. [27,28] Different ML algorithms cater to various problem types and can be classified into three major groups: supervised, unsupervised, and reinforcement learning. [29] Supervised ML algorithms are trained using labeled data, while unsupervised algorithms operate on unlabeled data. [27] The reinforcement learning method employs feedback-based learning, where a learning agent receives rewards or penalties based on task performance. [30] Additionally, ML can adopt a semi-supervised approach, training on a mix of labeled and unlabeled data. [31] Currently, ML methods have transitioned from purely theoretical constructs to practical applications. Although ML techniques effectively tackle complex tasks, they face several challenges, including the need for large amounts of high-quality data, the complexity of algorithms, and difficulties in interpreting results. [32] On the other hand, the QSPKR is a statistical method that quantitatively correlates chemical structures with their pharmacokinetic properties. [33] It is analogous to the quantitative structure-activity relationship (QSAR), which correlates chemical structures with activity rather than the PK parameters of the chemicals. [34] The QSPKR method can also be viewed as a supervised ML algorithm since it relies on labeled training data. In this method, various models are developed by analyzing significant amounts of data and subsequently used to predict PK parameters for unknown chemicals.

Previously, various attempts have been made to model drugs' tissue-to-plasma partition coefficient (Kp) values. [16,17,35–38] Graham et al. have shown that *in silico* models obtained from the literature can accurately predict the $K_p$ values of drugs in different tissues in rats. [17] Mayumi et al. developed a PBPK model for predicting $K_p$ values using the *in vitro* distribution volume at steady state ($V_{ss,\ in\ vitro}$), with $V_{ss,\ in\ vitro}$ data derived from protein binding data in rats. [35] Another PBPK model was created by Yun et al. to predict $K_p$ values of drugs using a decision tree-based classification algorithm. [16] Handa et al. reported two distinct *in silico* models for predicting $K_p$ values of drugs. In the first study, the authors introduced a 2D-QSAR model, developing random forest-based predictive models using physicochemical properties alongside in vitro experimental data. [36] In the second study, they utilized a similar modeling approach, developing ML models using various experimental parameters. [37] Y. E. Yun and A. N. Edginton developed a regression model employing a multiple stepwise regression method for predicting $K_p$ values of drugs, with the rat volume of distribution at steady state ($V_{ss}$) as an explanatory variable alongside other physicochemical parameters. [38]

In the present study, we have developed various machine learning (ML) based quantitative structure-pharmacokinetic relationship (QSPKR) models for predicting the tissue-to-plasma partition coefficient ($K_p$) of pharmaceuticals. Initially, the models were constructed utilizing only the structural and physicochemical properties of the drugs, demonstrating high quality and predictive capability for the target variable. The predictivity of these models was subsequently enhanced by integrating $K_p$ values from other tissues as an explanatory variable. In both cases, linear models like ridge regression and linear support vector machine show better results compared to the non-linear models like random forest. Here, we have also performed a screening utilizing an external dataset collected from the SIDER database, which indicates that the compounds with higher tissue partitioning are more prone to show toxicity to that particular tissue. Finally, we created a Python-based tool to facilitate predicting $K_p$ values for drugs when experimental data are unavailable.

## 2. Materials and methods

### 2.1. Data collection

The experimental data for this work were collected from the previous study of Handa et al. [37] The collected data contain the tissue-to-plasma partition coefficient ($K_p$) for various drugs across different body tissues, along with SMILES strings of the drugs. A total of 119 drug compounds were collected, which have $K_p$ values for the total 11 tissues, like the brain, kidney, lungs, etc. However, most of the data points do not have $K_p$ values for all the tissues. Among the collected 119 compounds, 69 compounds have a $K_p$ value for adipose tissue, 42 compounds have a $K_p$ value for bone, 89 compounds have a $K_p$ value for brain, 66 compounds have a $K_p$ value for gut, 94 compounds have a $K_p$ value for heart, 92 compounds have a $K_p$ value for kidney, 85 compounds have a $K_p$ value for liver, 93 compounds have a $K_p$ value for lung, 104 compounds have a $K_p$ value for muscle, 64 compounds have a $K_p$ value for skin, and 34 compounds have a $K_p$ value for spleen. Initially, we used these data separately for model development, and later, data from other tissues were also incorporated for the model construction. The details regarding the collected data are presented in Table S1 of Supplementary Material 2 **(SM2)**.

### 2.2. Molecular structure generation and energy minimization

The collected SMILES strings of the drugs were used to generate the molecular structure using Marvin Sketch software (https://chemaxon.com/marvin). The transformations, like the addition of explicit hydrogen, aromatic conversion, and 3D cleaning, were performed during the molecular structure generation, which was then saved in.mol file format.

The generated molecular structures were subjected to energy minimization, which is essential for calculating 3D descriptors. The collected data includes various stereoisomers, such as the R and S forms of a drug. To conduct modeling without excluding these compounds, we utilized 3D descriptors along with other types of descriptors. The Merck Molecular 2 (MM2) force field was used for energy minimization, using the

ChemDraw 3D software (https://revvitysignals.com/products/research/chemdraw). The energy-minimized molecular structures were also saved in a.mol file, which was then merged into a single.sdf file using OpenBabel software. [39] This .sdf file was then used to calculate the molecular descriptors.

### 2.3. Descriptor calculation and data preprocessing

Molecular descriptors are the numerical representation of a molecule, which helps to encode the structural fragment or physicochemical properties. [40] However, not all information contained within the molecule, but partly, can be extracted from the experimental procedure. In recent times, more focus has been given to the theoretical approaches to encode molecular information. The encoded molecular information present in the descriptors is used to establish a quantitative relationship between molecular structure and biological activity. [41] The descriptors can be classified into different groups, like 0D to 7D, based on their dimensionality. With the increase of dimensionality, the computational complexity also increases, which makes the descriptor calculation process more computationally exhaustive and less reproducible. [34] In this study, we have used alvaDesc software for the calculation of the molecular descriptor. [42] The energy-minimized file was used as input for the computation of descriptors. Here, a total of 5799 descriptors of different classes were calculated initially.

The descriptor matrix contains noise, such as missing values, string values, and inter-correlated descriptors, that must be removed before modeling; otherwise, it may affect the model's performance. We have removed the descriptors that contain a missing value, a string value, descriptors with low variance, and descriptors with high inter-correlation value. Here, we have used a 0.1 variance cutoff and a 0.9 inter-correlation cutoff for the removal of the descriptors, using the tool DataPreTreatmentGUI (available from: https://teqip.jdvu.ac.in/QSAR_Tools/).

### 2.4. Dataset division

In the QSPKR modelling, a model needs to be statistically evaluated properly to determine its further application. An external set is required to evaluate a QSPKR model, but an external set is not always available to evaluate a model. In such a situation, the original collected dataset is divided into two parts, the training set and the test set, where the training set is used to construct the model, and the test set is used to validate the model. [43] The division is carried out in such a manner that both the training and test sets should exhibit a similar type of distribution in the chemical space, and there should be a significant amount of similarity between them. [44] The training-test division can occur in various ratios (the most common divisions are 7:3, 3:1, and 4:1), depending on the number of data points in the dataset. A smaller amount of data in the training set can restrict the training process of the model, potentially resulting in unreliable predictions for unknown chemicals. Conversely, a smaller amount of data in the test set can constrain the validation process. [45]

In this study, division was carried out in a 4:1 ratio using three distinct methods with varying seeding techniques. The methods employed for division were Kennard-Stone (KS), [46] Euclidean Distance-based (ED), and the activity-sorted (AS) method, where the first two approaches divided the data sets based on the descriptor matrix, while the last method utilized the response variable for dataset division. [34] The optimal division was then chosen based on various validation metrics calculated from the test sets.

### 2.5. Feature selection

The feature selection step involves choosing a subset of the input features that can accurately define the chemical space of the current dataset. The calculated descriptor matrix contains various types of noise and redundant information that must be removed before model development; otherwise, it may negatively impact model performance. In this study, we employed multi-layered feature selection algorithms to extract the key features for model construction. We utilized stepwise multiple linear regression (SMLR), [47] genetic algorithms (GA), [48] and best subset selection (BSS) to reduce the feature space. The initial feature pool is reduced by the SMLR and GA method, and features obtained from these methods were used as input for the BSS method. The BSS method tries to develop the model with every possible combination of the input features, which makes this process computationally exhaustive when the number of features is increased. So, to avoid this problem initially, we used the SMLR and GA methods to reduce the number of features. In the SMLR method, MLR models were developed by sequentially adding or removing features based on the "stepping criteria" ($F$). The $F$-value is used to decide on the inclusion or removal of features; in this study, we used $F$ to enter = 4.0 and $F$ to remove = 3.9 for performing SMLR. [34] The underlying principle of the GA method was derived from biological evolution, where, under certain environmental conditions, the species (feature) with the highest fitness score is selected for the next generation. [48] Roger and Hopfinger first introduced this method to develop QSAR models. [49]

### 2.6. QSPKR model development

Once the optimal subset of features was found, the models were then trained using different types of ML algorithms. In this study, we have developed ML-based QSPKR models using eight different types of regression-based supervised ML algorithms. The developed models are random forest (RF), [50] gradient boost (GB), [51] adaboost (ADB), [52] extreme gradient boost (XGB), [53] support vector machine (SVM) (or nonlinear SVM), [54] linear support vector machine (LSVM), [54] ridge regression (RR), [55] and partial least squares (PLS) [56]. We employed these modelling approaches because they are the most commonly used modeling approaches in scientific research. Here, Deep learning (DL) algorithms were not considered for modelling, as these require a large amount of high-quality data to develop reliable models. [57]

The RF, GB, ADB, and XGB are tree-based ensemble learning methods that use multiple decision trees (DTs) to generate the final predictions. RF is an ensemble learning approach where multiple DT models are built in parallel and trained on bootstrap datasets generated from the original dataset. [50] The GB and ADB models are also ensemble learning methods, but they construct DT models sequentially rather than in parallel, with each DT correcting its predecessor. The GB algorithm works by fitting the residual error produced by the previous DT to the current DT. [51] Conversely, the ADB model assigns more weight to training instances that were mispredicted by its predecessor. [52] XGB is an extended version of the GB algorithm that enhances performance in terms of modeling speed and predictive accuracy. [53] The SVM is a supervised machine learning algorithm used for both classification and regression tasks. Its main principle is to identify a decision boundary that effectively separates data points of different classes within an n-dimensional feature space. SVMs can be divided into two types: linear and nonlinear. Linear SVMs are used when data points can be separated by a linear decision boundary or hyperplane. Nonlinear SVMs (or simply SVMs) are applied when a linear boundary cannot separate the classes, using different kernel functions to project data points into an n-dimensional feature space. [54] The RR is a regularized version of linear regression that reduces overfitting by adding a penalty term to the objective function during training. [55] The PLS is a variation of multiple linear regression (MLR), used for modeling collinear, correlated, and noisy data. [56]

Here, the model development and validation were performed using standardized training and test sets, where standardization was performed using the mean and standard deviation of the training set, using the formula shown in Eq. 1.

$$X' = \frac{X - \mu}{\sigma} \quad (1)$$

In this context, $X'$ and $X$ represent standardized and non-standardized descriptor/response values, respectively, while μ and σ denote the mean and standard deviation of the corresponding column of the training set.

In this study, we initially developed models for all the tissue-to-plasma partition coefficients separately, using only structural and physicochemical descriptors. Then, the developed models were further refined using the Kp from the other tissue as an explanatory variable. All the models were developed with optimized hyperparameter settings, where optimization was performed by 5-fold cross-validation with "neg_mean_absolute_error" as an objective function. The model development and optimization were performed using different functions of the Python-based module scikit-learn.

### 2.7. Statistical evaluation of the developed models

As mentioned above, a QSPKR model must be properly evaluated to determine its further application. In this study, we have calculated three distinct types of metrics: quality, internal, and external validation metrics, for assessing the developed models. The quality of a developed model is indicated by metrics such as the determination coefficient ($R^2$), Mean Absolute Error of the training set ($MAE_{train}$), and Root Mean Squared Error of Calibration (RMSEC), which are computed solely from the training set. The $R^2$ metric can take any value below 1, with a threshold of 0.6 making it an acceptable model. The model's quality improves as the $R^2$ value approaches 1. The error metrics ($MAE_{train}$ and RMSEC) do not have specific threshold values but should be as small as possible. [43] Internal validation metrics are calculated using the training sets to assess the model's robustness and goodness of fit. We have calculated two metrics for internal validation: the leave-one-out correlation coefficient ($Q^2_{LOO}$) and the leave-one-out mean absolute error ($MAE_{LOO}$). The metric $Q^2_{LOO}$ can also take a value below 1, requiring a threshold of 0.5 for an acceptable model, while $MAE_{LOO}$ should be a minimum. The predictivity of a model towards unknown chemicals is evaluated through external validation metrics calculated using the test set. The external validation metrics include predictive $R^2$ ($Q^2_{F1}$), [58] $Q^2_{F2}$, [59] Mean Absolute Error of the test set ($MAE_{test}$), and Root Mean Squared Error of Prediction (RMSEP). The required threshold for $Q^2_{F1}$ and $Q^2_{F2}$ is 0.5 for a model to be considered good predictive, while $MAE_{test}$ and RMSEP should also be as small as possible. [34] The aforementioned metrics are represented mathematically from Eqs. 2 to 10.

$$R^2 = 1 - \frac{\sum (Y_{obs(training)} - Y_{calc(training)})^2}{\sum (Y_{obs(training)} - \overline{Y}_{obs(training)})^2} \quad (2)$$

$$MAE_{training} = \frac{\sum |Y_{obs(training)} - Y_{calc(training)}|}{n_{training}} \quad (3)$$

$$RMSEC = \sqrt{\frac{\sum (Y_{obs(training)} - Y_{calc(training)})^2}{n_{training}}} \quad (4)$$

$$Q^2_{LOO} = 1 - \frac{\sum (Y_{obs(training)} - Y_{loo\ predicted(training)})^2}{\sum (Y_{obs(training)} - \overline{Y}_{obs(training)})^2} \quad (5)$$

$$MAE_{LOO} = \frac{\sum |Y_{obs(training)} - Y_{loo\ predicted(training)}|}{n_{training}} \quad (6)$$

$$Q^2_{F1} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \overline{Y}_{obs(training)})^2} \quad (7)$$

$$Q^2_{F2} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \overline{Y}_{obs(test)})^2} \quad (8)$$

$$MAE_{test} = \frac{\sum |Y_{obs(test)} - Y_{pred(test)}|}{n_{test}} \quad (9)$$

$$RMSEP = \sqrt{\frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{n_{test}}} \quad (10)$$

Here, $Y_{obs(training)}$ = observed response value of the training set, $Y_{obs(test)}$ = observed response value of the test set, $Y_{calc(training)}$ = calculated response value of the training set, $Y_{pred(test)}$ = predicted response value of the test set, $\overline{Y}_{obs(training)}$ = mean of the observed response value of the training set, $\overline{Y}_{obs(test)}$ = mean of the observed response value of the test set, $Y_{loo\ predicted(training)}$ = leave-one-out predicted value of the training set, $n_{training}$ = number of compound in the training set, $n_{test}$ = number of compound in the test set.

### 2.8. SHAP feature importance analysis

In this study, we analyzed feature importance using the SHAP (SHapley Additive exPlanation) method. [60] The relative importance of the modeled features is determined by the Shapley value, developed by Lloyd Shapley based on the concept of fair distribution in cooperative game theory. [61] The Shapley value represents the mean marginal contribution, which is the difference in predicting a model obtained by including a feature versus excluding that feature. [62] SHAP feature importance can be accessed in two ways: locally (compound-specific) and globally (overall importance, calculated by taking the absolute mean of the local importance). Here, we calculated the importance of the global SHAP feature, represented in the form of a summary plot. In the summary plot, features are arranged along the Y-axis based on their importance (with the feature at the top having the highest importance), while the X-axis displays SHAP values for each feature across all data points. [22,63,64]

### 2.9. Applicability domain analysis

According to the third principle of the Organization for Economic Co-operation and Development (OECD) regarding QSAR modelling, a developed QSAR model should have a defined applicability domain (AD) to determine its generalizability. The AD of a QSAR model is a chemical space defined by the training set, indicating the limitations of the model in terms of structural and response domains. [65] A developed QSAR model generally covers a limited chemical space defined by its training set. If an unknown chemical falls within the AD of the developed model, we can only say that the molecule could satisfy the scope of the developed model and that the prediction made by the model is reliable. Conversely, if a molecule falls outside the AD, then the prediction may not be reliable. [66] In this study, we employed the leverage approach to determine the AD of the developed models. In this method, the leverage (h) value of a compound is obtained from the diagonal elements of the HAT matrix (H) (**Equations 11 and 12**). If a compound has a leverage (h) value higher than the threshold (h*) value (Eq. 13), it is considered an outside-AD compound; otherwise, it is considered an inside-AD compound. [67]

For the training set compounds: $H_{tr} = X_{tr}(X_{tr}{}^T X_{tr})^{-1} X_{tr}{}^T$ (11)

For the test set compounds: $H_{te} = X_{te}(X_{tr}{}^T X_{tr})^{-1} X_{te}{}^T$ (12)

$$h^* = \frac{3(p+1)}{n} \quad (13)$$

Here, $X_{tr}$ = descriptor matrix of the training set, $X_{te}$ = descriptor matrix of the test set, $p$ = number of descriptors, and $n$ = number of compounds.

### 2.10. Screening of the true external set

In this study, the developed final models were also used to screen

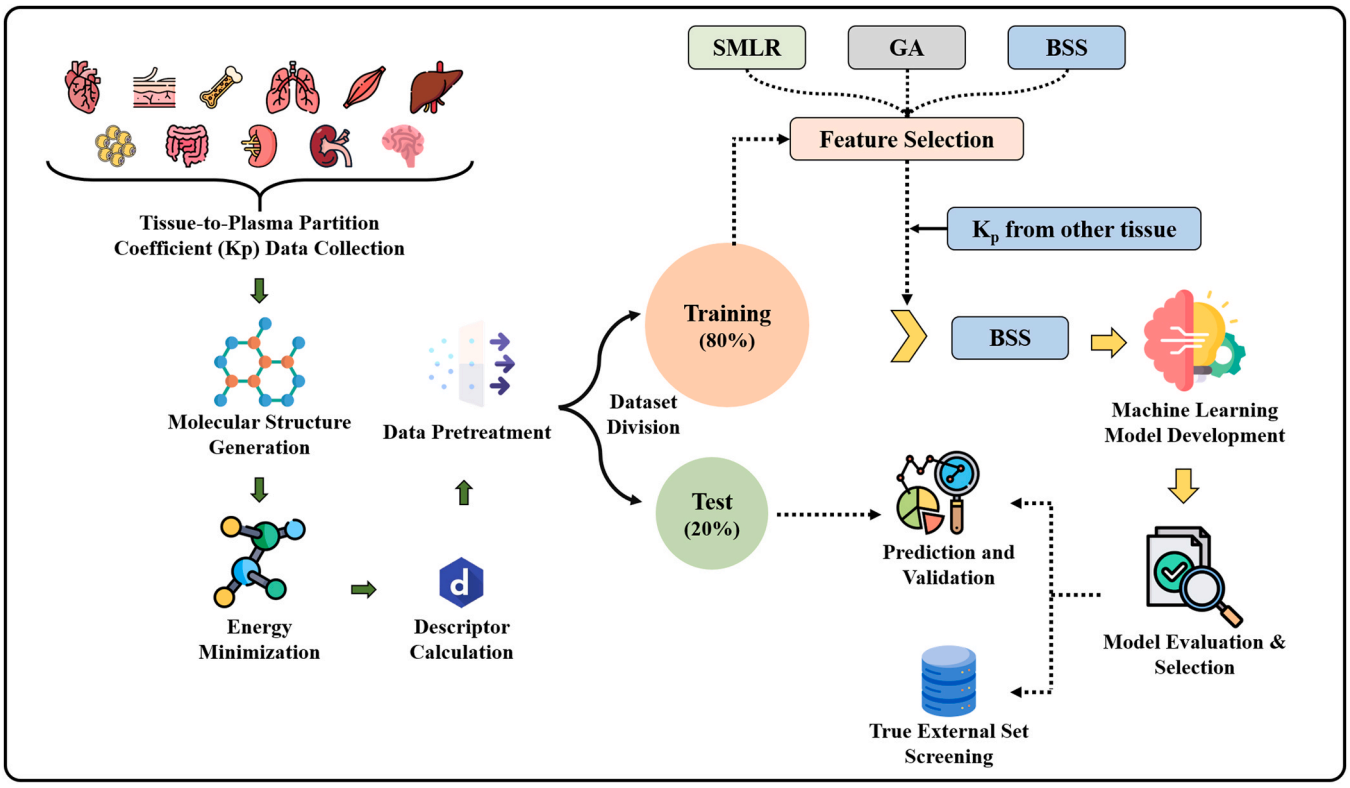



**Fig. 1.** Complete workflow of the present study.

true external datasets. We used the SIDER ('Side Effect Resource', http://sideeffects.embl.de) database for this purpose. [68] SIDER is a database of drugs and adverse drug reactions (ADRs) that is freely accessible for research. We obtained this data through the OCHEM database (https://ochem.eu/home/show.do). We collected data across eight categories: Skin and Subcutaneous Tissue Disorders, Renal and Urinary Disorders, Musculoskeletal or Tissue Disorders, Hepatobiliary Disorders, Cardiac Disorders, Nervous System Disorders, Gastrointestinal Disorders, and Respiratory Disorders. The dataset includes 3025 data points for respiratory disorders, while the other categories have a total of 1427 data points. Data curation was performed using KNIME software to remove salts, mixtures, ions, and inorganic compounds via the KNIME chemical workflow (available at: https://teqip.jdvu.ac.in/QSAR_Tools/). We followed the same cheminformatics workflow: Marvin Sketch for structure generation, ChemDraw 3D for energy minimization, and alvaDesc for descriptor calculation. Some data points yielded string values during descriptor calculation and were removed, resulting in a total of 2455 respiratory disorder data points and 1208 for the other categories. Our final models were used to predict the compounds, and the AD status was verified using the leverage method. All details about the data are presented in Table S2 of Supplementary Material 2 **(SM2)**.

The complete workflow of this study is represented in Fig. 1.

## 3. Results and discussion

In this study, we focused on the development of various ML-based QSPKR models aimed at predicting the tissue-to-plasma partition coefficient ($K_p$) across distinct tissues. Here, initially, we have developed multiple ML models utilizing properties derived only from the molecular structure of the drugs. Then, to enhance the accuracy and predictive power of these models, we subsequently incorporated $K_p$ values from the other tissues as independent variables, along with the structural and physicochemical properties of the drugs. This process allowed us to refine the models further, improving their predictive performance in estimating tissue distribution of the drugs.

### 3.1. Machine Learning model development

In this study, we initially developed ML models using structural and physicochemical descriptors calculated with alvaDesc software. The models were developed separately for eleven different tissues. The datasets for modeling were divided into training and test sets through various rational dataset division methods, with the training sets utilized for model construction and the test sets for model validation. The optimal training-test division was chosen based on external validation metrics. For all datasets, the division employed a 4:1 ratio, and in most instances, the methods based on Euclidean Distance (ED) and Activity Sorted (AS) yielded the best performance. Since the modeled dataset included different stereoisomers, distinct 3D descriptors were computed to encode these properties. As molecules must be optimized prior to calculating the 3D descriptors, we optimized the molecules using the MM2 force field. Initially, 5773 descriptors were calculated, which were then reduced to a modelable feature space using various feature selection techniques such as SMLR, GA, and BSS.

The eight types of ML models were developed for each dataset using the most significant descriptors identified through various feature selection techniques. Prior to model development, the datasets were scaled using the mean and standard deviation of the corresponding training sets. The best-performing models for each dataset, along with their optimized hyperparameter settings, are presented in Table 1. These results demonstrate the quality of the models, as well as their internal and external validation. As shown in the table, linear models, specifically **Ridge Regression (RR)** and **Linear Support Vector Regression (LSVR)**, yielded the best performance across all datasets (models consisting of physicochemical descriptors only). Each model achieved a quality metric, $R^2$, greater than 0.8 (with $R^2$ values ranging from 0.83 to 0.95) and a low training mean absolute error ($MAE_{train}$) ranging from 0.19 to 0.31, indicating high model quality. Additionally, the Leave-

**Table 1**
Results for the developed machine learning models, using structural and physicochemical properties.[a]

| Tissue | Division method | $n_{train}$ | $n_{test}$ | Descriptor Type | Algorithm | $n_{des}$ | $R^2$ | $MAE_{train}$ | RMSEC | $Q^2_{LOO}$ | $MAE_{LOO}$ | $Q^2_{F1}$ | $Q^2_{F2}$ | $MAE_{test}$ | RMSEP | Hyperparameters |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Adipose | AS−4 | 56 | 13 | Physicochemical | RR | 8 | 0.89 | 0.24 | 0.32 | 0.86 | 0.29 | 0.71 | 0.70 | 0.34 | 0.45 | alpha= 1.0 |
| | | | | Physicochemical + $K_p$ from other tissues | RR | 8 | 0.89 | 0.24 | 0.33 | 0.85 | 0.29 | 0.79 | 0.78 | 0.30 | 0.39 | alpha= 5.0 |
| Bone | AS−1 | 34 | 8 | Physicochemical | LSVR | 5 | 0.83 | 0.27 | 0.41 | 0.76 | 0.36 | 0.79 | 0.76 | 0.46 | 0.51 | C= 2.0 |
| | | | | Physicochemical + $K_p$ from other tissues | LSVR | 5 | 0.87 | 0.26 | 0.36 | 0.81 | 0.34 | 0.85 | 0.83 | 0.35 | 0.43 | C= 1.0 |
| Brain | ED−4 | 72 | 17 | Physicochemical | RR | 11 | 0.89 | 0.24 | 0.32 | 0.84 | 0.29 | 0.81 | 0.80 | 0.34 | 0.46 | alpha= 1.0 |
| | | | | Physicochemical + $K_p$ from other tissues | LSVR | 11 | 0.89 | 0.21 | 0.32 | 0.87 | 0.27 | 0.82 | 0.82 | 0.34 | 0.44 | C= 2.0 |
| Gut | ED−1 | 53 | 13 | Physicochemical | RR | 9 | 0.92 | 0.23 | 0.28 | 0.88 | 0.27 | 0.93 | 0.92 | 0.22 | 0.26 | alpha= 5.0 |
| | | | | Physicochemical + $K_p$ from other tissues | RR | 7 | 0.91 | 0.22 | 0.29 | 0.88 | 0.26 | 0.95 | 0.94 | 0.20 | 0.22 | alpha= 1.0 |
| Heart | ED−3 | 76 | 18 | Physicochemical | RR | 10 | 0.88 | 0.28 | 0.35 | 0.84 | 0.33 | 0.91 | 0.90 | 0.26 | 0.29 | alpha= 1.0 |
| | | | | Physicochemical + $K_p$ from other tissues | RR | 7 | 0.92 | 0.22 | 0.28 | 0.89 | 0.24 | 0.95 | 0.95 | 0.16 | 0.21 | alpha= 1.0 |
| Kidney | ED−2 | 74 | 18 | Physicochemical | RR | 11 | 0.88 | 0.27 | 0.34 | 0.84 | 0.32 | 0.82 | 0.81 | 0.43 | 0.56 | alpha= 0.5 |
| | | | | Physicochemical + $K_p$ from other tissues | RR | 8 | 0.87 | 0.27 | 0.35 | 0.84 | 0.30 | 0.91 | 0.91 | 0.25 | 0.39 | alpha= 5.0 |
| Liver | AS−1 | 68 | 17 | Physicochemical | LSVR | 12 | 0.85 | 0.24 | 0.38 | 0.81 | 0.32 | 0.81 | 0.81 | 0.34 | 0.43 | C= 1.0 |
| | | | | Physicochemical + $K_p$ from other tissues | LSVR | 11 | 0.85 | 0.24 | 0.38 | 0.81 | 0.30 | 0.90 | 0.90 | 0.23 | 0.31 | C= 3.0 |
| Lung | ED−3 | 75 | 18 | Physicochemical | RR | 10 | 0.93 | 0.19 | 0.26 | 0.91 | 0.21 | 0.90 | 0.90 | 0.25 | 0.33 | alpha= 5.0 |
| | | | | Physicochemical + $K_p$ from other tissues | RR | 7 | 0.95 | 0.18 | 0.23 | 0.94 | 0.19 | 0.94 | 0.94 | 0.20 | 0.27 | alpha= 5.0 |
| Muscle | KS | 84 | 20 | Physicochemical | RR | 9 | 0.85 | 0.31 | 0.39 | 0.79 | 0.36 | 0.82 | 0.77 | 0.34 | 0.46 | alpha= 1.0 |
| | | | | Physicochemical + $K_p$ from other tissues | RR | 7 | 0.88 | 0.26 | 0.35 | 0.85 | 0.29 | 0.91 | 0.89 | 0.23 | 0.32 | alpha= 1.0 |
| Skin | AS−1 | 52 | 12 | Physicochemical | RR | 8 | 0.85 | 0.29 | 0.38 | 0.79 | 0.34 | 0.72 | 0.70 | 0.35 | 0.55 | alpha= 0.5 |
| | | | | Physicochemical + $K_p$ from other tissues | RR | 6 | 0.90 | 0.23 | 0.31 | 0.86 | 0.27 | 0.94 | 0.94 | 0.17 | 0.25 | alpha= 0.5 |
| Spleen | AS−3 | 28 | 6 | Physicochemical | LSVR | 5 | 0.95 | 0.14 | 0.22 | 0.94 | 0.17 | 0.89 | 0.87 | 0.22 | 0.26 | C= 2.0 |
| | | | | Physicochemical + $K_p$ from other tissues | PLS | 4 | 0.96 | 0.15 | 0.19 | 0.95 | 0.18 | 0.89 | 0.87 | 0.23 | 0.26 | LVs= 2 |

[a] RR = Ridge Regression, LSVR = Linear Support Vector Regression, PLS = Partial Least Squares, AS = Activity Sorted, ED = Euclidean Distance, KS = Kennard Stone, $n_{train}$ = number of compounds in the training set, $n_{test}$ = number of compounds in the test set, $n_{des}$ = number of descriptors.

One-Out cross-validation metric $Q^2_{LOO}$ ranged from 0.76 to 0.94, while the external validation metrics $Q^2_{F1}$ (ranging from 0.71 to 0.93) and $Q^2_{F2}$ (ranging from 0.70 to 0.92) further reflect the predictability of the models. Here, all the ML models were developed with optimized hyperparameter settings, where optimization was conducted using the GridSearchCV function from Scikit-learn, employing negative mean absolute error as the objective function along with five-fold cross-validation. Here, we have generated scatter plots (Figure S1 **of** Supplementary Material 1) for all the developed models by plotting observed $K_p$ along the X-axis and predicted $K_p$ along the Y-axis. From this plot, we can see that all the models show a good positive correlation between observed and predicted $K_p$ values. We have also generated a validation curve (Figure S2 **of** Supplementary Material 1) for all the datasets to represent the optimal hyperparameter selection for the developed models.

The developed models were further refined by incorporating $K_p$ values from other tissues as explanatory variables. To utilize $K_p$ values from other tissues, we employed predicted values when experimental values were unavailable. Here, previously modeled descriptors, along with $K_p$ values from other tissues, were used as input for BSS feature selection to identify the most significant features for model development. Similarly, eight different types of ML models were developed with scaled training and test sets. The results of these models are also presented in Table 1. The findings indicate a significant improvement in model performance regarding predictability, achieved with the same or fewer features. For most models, the $R^2$ values remain nearly unchanged, but the external validation metrics, $Q^2_{F1}$ (ranging from 0.79 to 0.95) and $Q^2_{F2}$ (ranging from 0.78 to 0.95), show considerable improvement. Furthermore, linear models, specifically **RR**, **LSVR**, and **partial least squares (PLS)**, demonstrated superior performance compared to other models. The scatter plots (Fig. 2) for this case were also generated, showing a good positive linear correlation between the observed and predicted Kp values. Furthermore, models were developed with optimized hyperparameter settings, and the optimal hyperparameter selection is illustrated by the validation curve shown in Fig. 3.

### 3.2. Feature importance analysis

In this study, we conducted a feature importance analysis using the SHapley Additive Explanations (SHAP) method to determine the relative importance of the modeled features with respect to the response parameters of the corresponding data sets. The feature importance is based solely on the training sets of the final model, which include both types of descriptors (physicochemical and $K_p$ values from other tissues). In this context, SHAP feature importance is presented as a summary plot, where SHAP values for each observation are plotted along the X-axis, and features are organized along the Y-axis according to their importance (with the feature at the top indicating the highest importance). The SHAP summary plots for all the developed models are represented in Figs. 4**a to** 4**d and** Figure S3a to S3g **of** Supplementary Material 1 **(SM1)**. Here, Table 2 provides a comprehensive exploration of how structural and physicochemical descriptors influence the response parameters.

From Table 2, we can observe that fifty distinct 2D and 3D structural and physicochemical descriptors were found to be significant for modeling the $K_p$. Most of the 2D descriptors belong to the categories of 2D atom pair, atom type E-state descriptors, and constitutional indices. Various drug-like indices, such as Depressant-80 (adipose), Psychotic-80 (brain and lung), and Neoplastic-50 (muscle), were identified as significant for model development. All of these drug-like indices exhibit a positive contribution to the specified target variable. The 3D descriptors, such as pharmacophore descriptors (CATS2D and CATS3D) and 3D-MoRSE descriptors, demonstrate higher significance for model development.

Here, the performance of all developed models was significantly enhanced by incorporating the $K_p$ values from other tissues. The

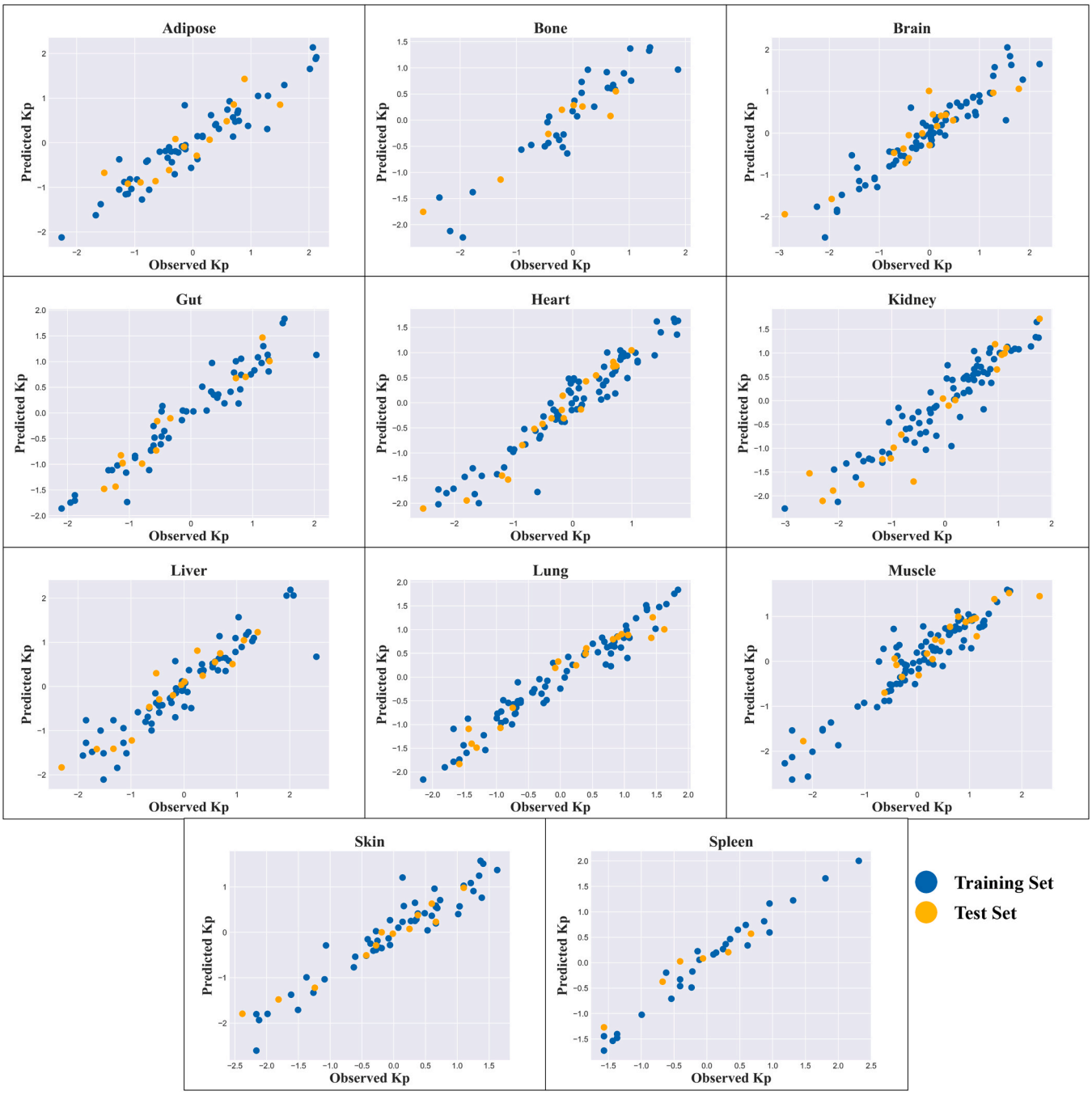



**Fig. 2.** Scatter plots for the machine learning models developed using structural and physicochemical descriptors along with tissue-to-plasma partition coefficient ($K_p$) from other tissues.

direction of contribution (positive or negative, based on the SHAP summary plot) of the $K_p$ values from other tissues to the target tissue is also illustrated in a heatmap-like plot (Figure S4 of Supplementary Material 1). Most of the $K_p$ values indicate a positive contribution to the target tissue. The $K_p$ values of the heart, liver, and lung appear in most models, whereas the $K_p$ values of the spleen are not present in any of the models. The impact of the $K_p$ values from other tissues on the target tissue is discussed below:

*Heart*: The heart is found to be influential in most of the developed models, showing a positive contribution to the target tissues like bone, brain, kidney, lung, muscle, and spleen. The positive contribution can be explained based on the dynamic distribution of the drug to the body, which makes the drug available for partitioning in the other tissues. However, high cardiac Kp (indicating significant drug retention in heart tissue) may reduce systemic drug availability, indirectly lowering exposure to the other tissues.

*Liver*: The liver $K_p$ values positively contribute to the $K_p$ value of the adipose, gut, heart, kidney, and skin, which may be attributed to different interconnected PK, metabolic, and inflammatory mechanisms. A higher liver $K_p$ value reduces the plasma drug concentration by metabolizing the drug. This reduction in plasma concentration increases the apparent $K_p$ values in other tissues (tissue concentration (unchanged)/plasma concentration (decrease)). The different disease conditions of the liver (e.g., fatty liver) release pro-inflammatory cytokines, which may enhance the tissue permeability or transporter expression, promoting drug accumulation in other tissues. [76,77] If different

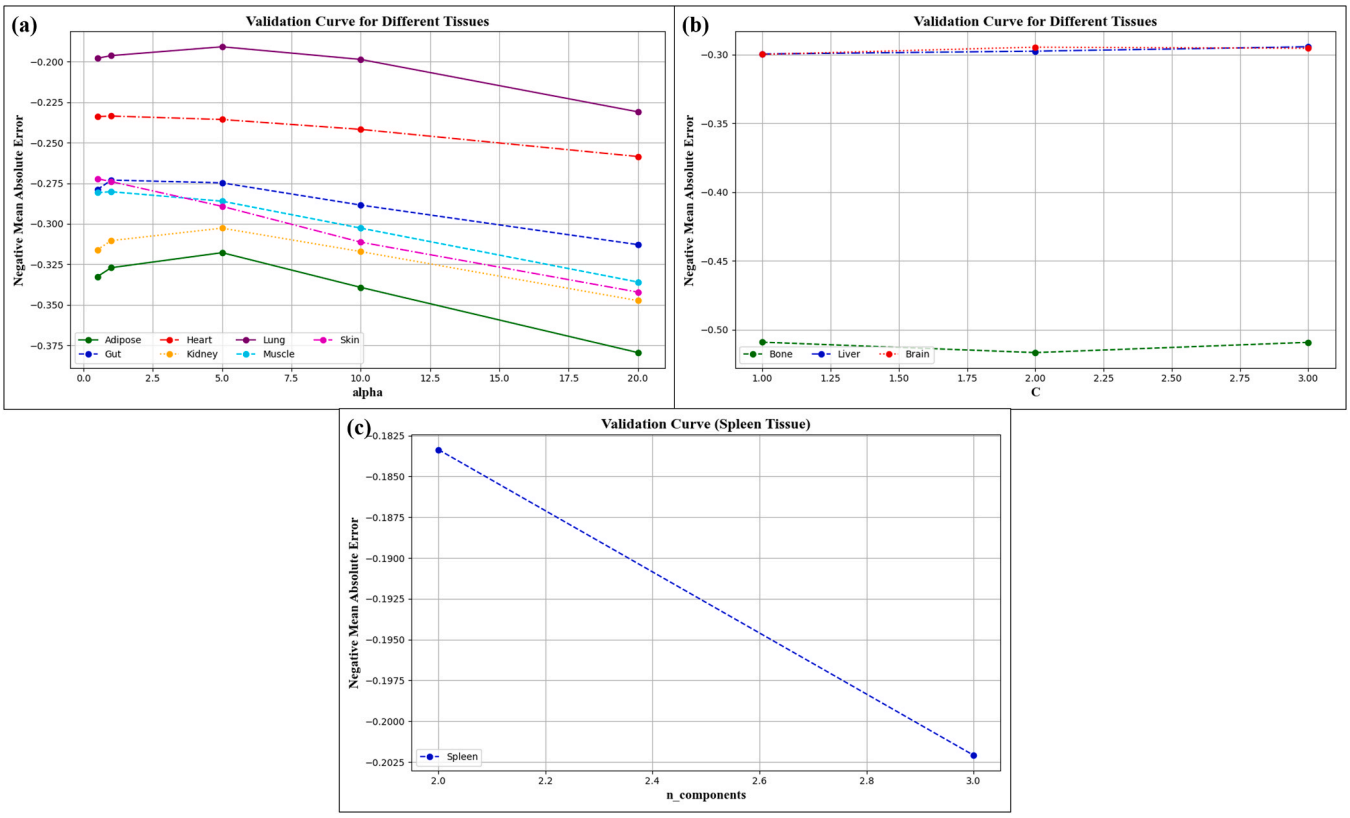



**Fig. 3.** Validation curves of the developed ML models (Physicochemical + Kp from other tissues) for different datasets. (a) Validation curve for the ridge regression models. (b) Validation curve for the linear support vector machine models (c) Validation curve for the partial least squares model (spleen tissue).

tissues had a similar type of active transporter for a drug, then these tissues could show similar distribution for a drug and be influenced by each other.

*Brain and Gut*: The brain shows a negative influence on the partitioning of a drug in the gut tissue, but the gut positively influences the partitioning of a drug in the brain tissue. The gut tissue influences the brain's $K_p$ indirectly through enterohepatic circulation, where bile-excreted drug is reabsorbed in the intestine and introduced to the systemic circulation. [78,79] Here, the brain's Kp value contributes negatively to the gut's partitioning of a drug by sequestering the drug, reducing systemic availability.

*Adipose tissue*: The adipose tissue positively influences drug partitioning in the skin while negatively impacting partitioning in the kidney. The skin contains a significant amount of fat in the subcutaneous layer, and skin-associated adipocytes actively interact with skin cells to maintain skin health. [80] The positive influence of adipose tissue on drug partitioning in the skin indicates a greater drug affinity for the lipophilic environment. In conditions like obesity, the amount of fat in the subcutaneous layer increases, which enhances drug partitioning in the skin. [81] The kidney, being a highly vascular organ, has a low fat content, leading various lipophilic drugs to accumulate in fat-rich tissue. The partitioning of a drug in fat-rich organs decreases its availability in systemic circulation and diminishes its partitioning in other tissues.

The other organs, like the lungs, skin, bone, muscle, and kidneys, also influence (positively or negatively) the partitioning of a drug in other tissues. The tissues that exhibit a similar physiological environment (e.g., water content, protein binding, lipid content) indicate that the factors determining drug partitioning in one tissue may also be relevant to other tissues. Therefore, in such situations, these tissues display similar drug distribution properties. A drug shows comparable partitioning if it shares similar characteristics (e.g., lipophilicity, perfusion rate) with target tissues. However, if a tissue demonstrates distinct composition, then the partitioning of a drug in that particular tissue will increase, while free drug in systemic circulation and partitioning in other tissues will be reduced.

### 3.3. Applicability domain analysis

In this study, we used the leverage approach for the applicability domain (AD) analysis of the final developed models, where leverage values are computed using only the structural domain (descriptor matrix) of the training and test sets. The finding of this analysis is represented in the form of a scatter plot, where leverage values are plotted along the Y-axis and compound IDs are plotted along the X-axis. The threshold value (h*) is represented by a horizontal line parallel to the X-axis, indicating that any observation above the line is considered the outside-AD compound. The AD plot of the developed models is represented in Figs. 5**a** to 5**d** and Figure S5a to S5g of Supplementary Material 1 **(SM1)**. From these plots, we can see that only a few number of compounds fall outside of the AD of the developed models.

### 3.4. Screening of the true external set

In this study, we utilized our final selected models to predict the partitioning of various compounds within a particular tissue, collected from the SIDER database. The SIDER database provides information on drugs and their associated adverse drug reactions (ADRs). The predicted values for these compounds, along with their AD status, are presented in Table S2 of Supplementary Material 2 **(SM2)**. A compound that exhibits a higher partitioning in a specific tissue is more likely to cause an adverse effect to that tissue. We analyzed the top fifty compounds that exhibited the highest tissue-to-plasma partition coefficients for a particular tissue (within the AD). Our findings indicate that, in most cases, the developed models predict that higher partitioning correlates with ADR for that tissue. This supports the effectiveness of our models in predicting the behavior of unknown compounds. In Table 3, we present

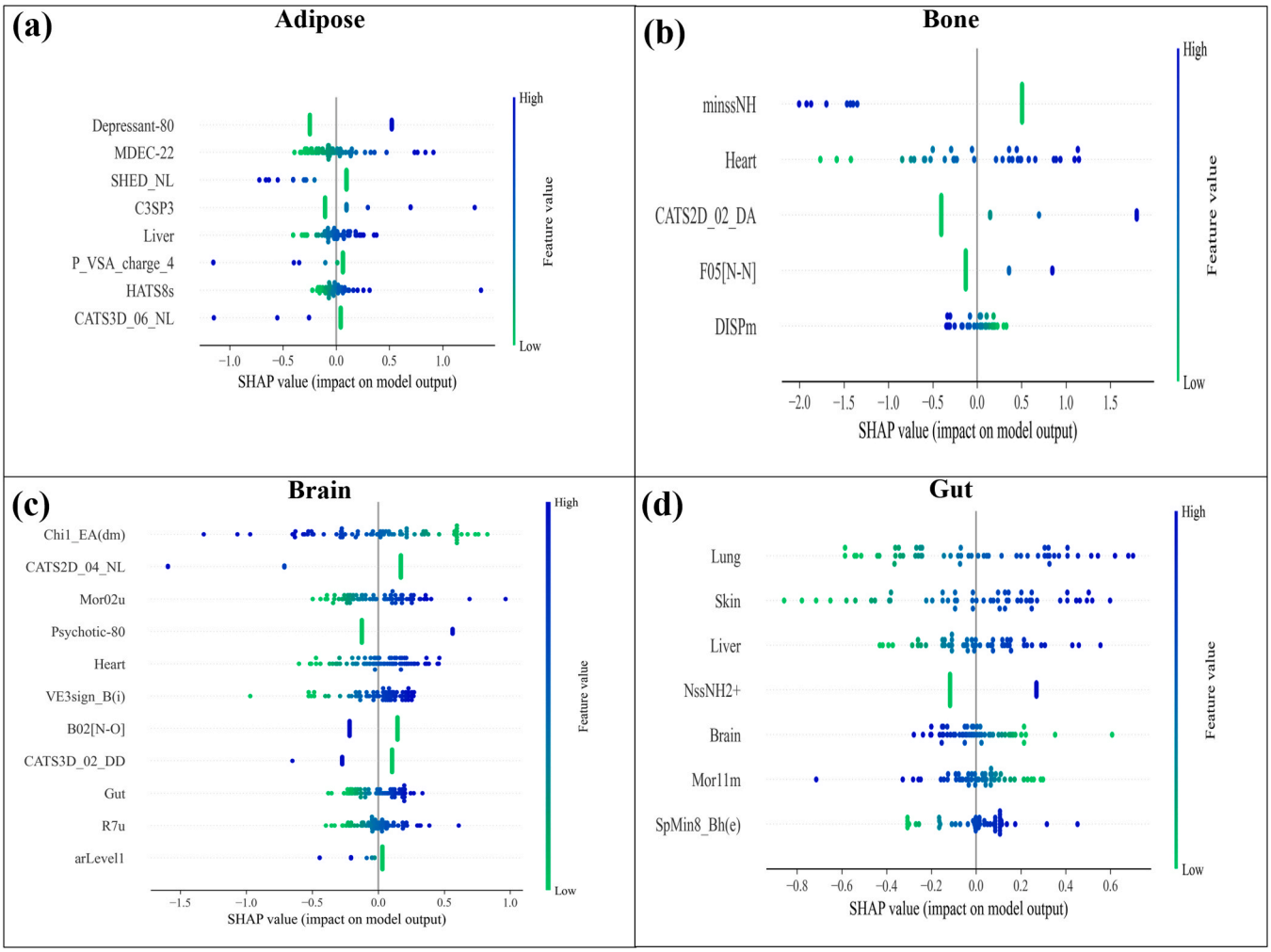



**Fig. 4.** SHAP summary plots. (a) Summary plot for the Adipose tissue. (b) Summary plot for Bone. (c) Summary plot for Brain. (d) Summary plot for Gut.

the top four compounds that have the highest predicted partition coefficient values (within the AD) and also demonstrate toxicity for the corresponding tissue.

### 3.5. Comparison with previous studies

Here, we conducted a comparative analysis with the previous studies performed by Handa et al. [36,37] The previous authors performed two different studies to develop predictive models for the $K_p$ of drugs. In the first study, they developed 2D-QSAR models using various in vitro parameters (e.g., LogP, fraction of ionization (fi), pKa, $f_{up}$, and ionization class (acid, base, neutral, etc.)) as well as different structural and physicochemical properties of the drugs. In this study, they did not split the dataset; instead, they created models with all the compounds and validated them using the cross-validation technique. In the second case, they developed predictive models for the $K_p$ values of drugs in various tissues, employing various in vitro parameters alongside the $K_p$ of other tissues as explanatory variables. In both studies, they utilized different machine learning algorithms for model development. Here, we compare these studies with our current research and present the findings in Table 4. From this table, we can see that our models perform better than previously developed models in terms of model quality and predictability. Another advantage of our model is that it can predict the $K_p$ of a drug without requiring any *in vitro* parameters. Here, prediction is performed by using a drug's structural and physicochemical properties, along with the $K_p$ of other organs as explanatory variables. If the $K_p$ values of the other organs are unavailable, they can be predicted from the initial models.

### 3.6. Development of the software

To enhance user convenience, we have developed a Python-based graphical user interface (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/kp-calculator) (Fig. 6) for predicting the $K_p$ values of drugs. This tool is available for free on the DTC Lab website. Users need to provide descriptor inputs in an.xlsx file format to make predictions. Additionally, the tool includes information about the AD status, indicating the reliability of predictions.

## 4. Conclusion

The tissue-to-plasma partition coefficient ($K_p$) is one of the important PK parameters that determine the transferability of a drug within a tissue. This PK parameter is also essential for assessing a drug's tissue-specific activity or toxicity. The experimental determination of the $K_p$ parameter requires a large number of animal experiments, regular sample collection, and analysis, which makes this process resource-intensive. Therefore, an alternative approach needs to be developed to determine the $K_p$ values of drugs that can be derived from easily accessible parameters. In this study, we developed machine learning (ML) - based quantitative structure-pharmacokinetic relationship (QSPKR) models for predicting the tissue-to-plasma partition coefficient of drugs. Here, the ML models were developed for eleven different tissues. Initially, the models were constructed using the structural and

**Table 2**
Structural and physicochemical descriptors present in the final model.

| Descriptor | Explanation | Class | Direction of contribution | Target Tissue |
|---|---|---|---|---|
| CATS3D_06_NL | It represents the CATS3D negative lipophilic bin of 06. The presence of negative lipophilic groups (O atom of hydroxyl or carboxyl group, etc.) in a molecule can negatively influence the partitioning of a molecule within a tissue. The negative lipophilic groups decrease the affinity for the non-polar environment, indicating a higher affinity for water (hydrophilic). | CATS 3D | (-) ve | **Adipose** |
| Depressant−80 | This descriptor represents the Ghose-Viswanadhan-Wendoloski antidepressant-like index at 80 %, first proposed by Ghose, Viswanadhan, and Wendoloski in 1999. [69] It was developed by analyzing the physicochemical properties of a specific set of drugs—namely, antidepressants. Four main properties are considered: lipophilicity (ALOGP), molar refractivity (AMR), molecular weight (MW), and the number of atoms (nAT). The term 80 % indicates the percentage of the distribution of these properties calculated from a reference set of known drugs. In other words, researchers analyze a dataset of known drugs and determine the ranges of molecular properties that cover 80 % of those drugs. This descriptor helps predict the drug-likeness of a chemical substance with similar physicochemical properties. The ranges of the physicochemical properties for the depressant−80 descriptor are: ALOGP, 1.4–4.9; AMR, 62–114; MW, 210–380; and nAT, 32–56. If a compound falls within all four property ranges, it is said to pass the index at 80 %, indicating a high likelihood of antidepressant-like potential. [69] Based on the above discussion, we can conclude that compounds with a Depressant−80 value are likely to exhibit greater lipophilicity, as they share common physicochemical properties with CNS-active drugs. This characteristic can enhance tissue partitioning, which is also supported by its positive contribution. | Drug-like indices | (+) ve | |
| C3SP3 | It represents the sp3 hybridized carbon atoms attached to three other carbon atoms in the molecule. This descriptor affects the partitioning of a drug in the tissue by increasing its lipophilicity. | Topological indices | (+) ve | |
| HATS8s | The descriptor HATS8s is a GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptor, calculated by the leverage-weighted autocorrelation of molecular graphs. The atomic weight and intrinsic state (*I*-state) are used for its calculation. | GETAWAY descriptors | (+) ve | |
| MDEC−22 | This is a molecular distance-edge vector, indicating the molecular distance edge between all secondary carbon atoms present in the molecule. | MDE descriptors | (+) ve | |
| P_VSA_charge_4 | It indicates the amount of van der Waals surface area (VSA) having the partial charges in bin 4. This descriptor contributes negatively, as it represents the charge distribution across the molecule, which limits the partition of a molecule within a tissue. | P_VSA-like descriptors | (-) ve | |
| SHED_NL | It represents the Shannon Entropy Descriptors of the negative lipophilic component of a molecule. This descriptor also shows effects similar to CATS3D_06_NL, which increase hydrophilicity and reduce tissue partitioning. | Pharmacophore descriptors | (-) ve | |
| minssNH | It represents the molecular fragment with the minimum electronic state of a single bond-single bond NH group present in a molecule. This is the most important descriptor for the bone tissue, which shows a negative contribution. This group may influence the overall polarity of a molecule, which can affect the partitioning between the aqueous plasma compartments and the organic interface of bone tissue. The local pH and mineral content in bone may alter the NH group's ionization state, potentially reducing its affinity for bone. [70] | Atom-type E-state indices | (-) ve | **Bone** |
| CATS2D_02_DA | It represents the count of the number of pairs of hydrogen bond donor-acceptor atoms that are separated by a topological distance of two bonds. The higher value of this descriptor indicates that the molecule has a higher possibility of forming a hydrogen bond with bone matrix, which affects the partitioning of a molecule. [71] | Pharmacophore descriptors | (+) ve | |
| DISPm | It represents the displacement value weighted by mass. | Geometrical descriptors | (-) ve | |
| F05[N-N] | It represents the frequency of two nitrogen atoms at a topological distance of five, which contributes positively to the response parameter. As the value of this descriptor increases, the number of nitrogen atoms within the molecule also increases, which may help form hydrogen bonds with the mineral surface area of the bone. [71] | 2D Atom Pairs | (+) ve | |
| B02[N-O] | Presence or absence of nitrogen and oxygen atoms at a topological distance of two. The presence of such atoms in a molecule leads to an increase in the polarity of the molecule, which in turn leads to an increase in the hydrophilicity of the molecule and a decrease in the partitioning of the drug within the brain. | 2D Atom Pairs | (-) ve | **Brain** |
| CATS2D_04_NL | Count the number of pairs of negatively charged-lipophilic atoms that are separated by a topological distance of four bonds. This descriptor may decrease the partitioning of a molecule into the brain by increasing its hydrophilicity. | Pharmacophore descriptors | (-) ve | |
| arLevel1 | It represents the number of neighbouring aromatic atoms of a molecule's chiral centre (level 1). | Chirality descriptors | (-) ve | |
| Mor02u | It refers to the signal or index number 02 in the MoRSE descriptor series, where u indicates that the descriptor is unweighted. | 3D-MoRSE descriptors | (+) ve | |
| Chi1_EA(dm) | This feature indicates the connectivity-like index of order 1 from the edge adjacency matrix, weighted by dipole moment. This descriptor has the highest importance towards the response parameters and contributes negatively to it by affecting the polarity of the molecule. | Edge adjacency indices | (-) ve | |
| VE3sign_B(i) | It represents logarithmic coefficient sum of the last eigenvector from the Burden matrix, weighted by ionization potential. | 2D matrix-based descriptors | (+) ve | |
| R7u | R autocorrelation of lag 7 / unweighted. | GETAWAY descriptors | (+) ve | |
| CATS3D_02_DD | It represents the number of hydrogen bond donor–donor pairs in bin 2. This descriptor contributes negatively by increasing the hydrogen bonding probability. | CATS 3D | (-) ve | |
| Psychotic−80 | It indicates the Ghose-Viswanadhan-Wendoloski antipsychotic-like index at 80 %. This descriptor also helps to determine drug-likeliness, similar to the Depressant−80, which is generated by analysing the dataset of antipsychotic drugs. The ranges of the physicochemical properties for the Psychotic−80 descriptor are: ALOGP, 2.3–5.2; AMR, 85–131; MW, 274–464; | Drug-like indices | (+) ve | **Brain, Lung** |

(*continued on next page*)

**Table 2** (*continued*)

| Descriptor | Explanation | Class | Direction of contribution | Target Tissue |
|---|---|---|---|---|
| | and nAT, 40–63. If a compound falls within all four property ranges, it is said to pass the index at 80 %, indicating a high likelihood of antipsychotic-like potential. [69] The compounds with Psychotic−80 values exhibit greater lipophilicity since they are also similar to CNS-active drugs. Consequently, these compounds have increased partitioning in lipid-rich tissues. | | | |
| SpMin8_Bh(e) | This descriptor represents the smallest eigenvalue n. 8 of the Burden matrix, weighted by Sanderson electronegativity. | Burden eigenvalues | (+) ve | **Gut** |
| Mor11m | It refers to the signal or index number 11 in the MoRSE descriptor series, where m indicates that the descriptor is weighted by atomic mass. | 3D-MoRSE descriptors | (-) ve | |
| NssNH2 + | This descriptor indicates the number of atom types of the single bond, single bond NH2 + group. | Atom-type E-state indices | (+) ve | |
| B06[C-N] | This descriptor represents the presence or absence of carbon and nitrogen atoms at a topological distance of six, which shows positive contribution to the response parameter. This feature highlights two important aspects of molecular structure. First, the presence of large fragments within a molecule enhances its lipophilicity. Second, the inclusion of electronegative atoms gives the molecule a slight polarity. These characteristics help the molecule remain water-soluble while still having an affinity for membranes. | 2D Atom Pairs | (+) ve | **Heart** |
| Mor04m | It refers to the signal or index number 04 in the MoRSE descriptor series, where m indicates that the descriptor is weighted by atomic mass. | 3D-MoRSE descriptors | (+) ve | |
| B02[O-O] | Presence or absence of two oxygen atoms at a topological distance of two. The presence of such electronegative atoms makes the molecule polar, enhancing its interaction with the aqueous environment and decreasing tissue partitioning. | 2D Atom Pairs | (-) ve | **Heart, Spleen** |
| MaxdssC | It represents the molecular fragment with the maximum electronic state of a carbon atom with a double bond, and two single bonds (>C=) present in a molecule. | Atom-type E-state indices | (+) ve | **Kidney** |
| Mor12m | It refers to the signal or index number 12 in the MoRSE descriptor series, where m indicates that the descriptor is weighted by atomic mass. | 3D-MoRSE descriptors | (-) ve | |
| nN+ | This descriptor indicates the number of positively charged nitrogen atoms in a molecule, which can exist as either NH+ or NH2 + in a cyclic ring system or as part of a side chain. It is most significant for kidney tissue and the second most significant for spleen tissue. In both types of tissue, an increase in this descriptor value correlates with an enhanced partitioning of the compound within the tissue. This effect may be attributed to interactions with the extracellular matrix. | Functional group count | (+) ve | **Kidney, Spleen** |
| CATS2D_02_AN | It represents the count of the number of pairs of hydrogen bond acceptors and negatively charged atoms that are separated by a topological distance of two bonds. | Pharmacophore descriptors | (-) ve | **Liver** |
| VE3sign_B(v) | This descriptor represents the logarithmic coefficient sum of the last eigenvector from the Burden matrix, weighted by van der Waals volume. | 2D matrix-based descriptors | (-) ve | |
| Mor02i | It refers to the signal or index number 02 in the MoRSE descriptor series, where i indicates that the descriptor is weighted by ionization potential. | 3D-MoRSE descriptors | (+) ve | |
| arLevel2 | It represents the number of neighbouring aromatic atoms of a molecule's chiral centre (level 2). | Chirality descriptors | (-) ve | |
| C−039 | It represents the fragment Ar-C(=X)-R, where Ar = aromatic groups, X = any electronegative atom (e.g., O, N, S, P, Se, halogens), and R = any group that is linked through carbon. | Atom-centred fragments | (+) ve | |
| CATS2D_07_PL | It indicates the number of pairs of positively charged lipophilic atoms that are separated by a topological distance of seven bonds. | Pharmacophore descriptors | (+) ve | |
| B06[N-O] | This descriptor represents the presence or absence of nitrogen and oxygen atoms at a topological distance of six. It helps to increase the hydrophilicity of the molecule by incorporating the electronegative atoms and decreasing the tissue partitioning. | 2D Atom Pairs | (-) ve | |
| CATS2D_09_LL | This descriptor refers to the count of the number of pairs of lipophilic-lipophilic atoms that are separated by a topological distance of nine bonds. | Pharmacophore descriptors | (-) ve | |
| CATS2D_04_AL | It indicates the count of the number of pairs of hydrogen bond acceptor-lipophilic atoms that are separated by a topological distance of four bonds. | Pharmacophore descriptors | (+) ve | |
| D/Dtr03 | It indicates the distance/detour ring index of order 3. | Ring descriptors | (+) ve | **Lung** |
| THSA | This descriptor represents the total hydrophobic surface area of the molecules. Therefore, a higher descriptor value (i.e., increased hydrophobic surface area) can improve a drug's partitioning into tissue. | SASA descriptors | (+) ve | |
| Mor17s | It refers to the signal or index number 17 in the MoRSE descriptor series, where s indicates that the I-state weights the descriptor. | 3D-MoRSE descriptors | (+) ve | |
| P_VSA_LogP_4 | It indicates the amount of van der Waals surface area (VSA) in the fourth LogP bin. | Constitutional indices | (-) ve | |
| totalcharge | This descriptor indicates the total charge present in the molecule. This is the second most important descriptor for the partitioning of the drugs within lungs and muscle; the least important for the skin tissue. This shows a positive contribution to the lung and muscle, while shows a negative contribution to the skin. This difference may be attributed to the composition of the tissues: muscle and lung tissue contain less fat and have a higher water content, whereas skin has a higher lipid content. [72-75] | Constitutional indices | (+) ve<br>(-) ve | **Lung, Muscle**<br>**Skin** |
| F02[C-S] | This descriptor represents the frequency of carbon and sulphur atoms at a topological distance of two. | 2D Atom Pairs | (-) ve | **Muscle** |
| F06[C-Cl] | This descriptor represents the frequency of carbon and chlorine atoms at a topological distance of six. | 2D Atom Pairs | (+) ve | |
| CATS2D_04_AN | It indicates the count of the number of pairs of hydrogen bond acceptors and negatively charged atoms that are separated by a topological distance of four bonds. This descriptor shows positive contribution which may be due to their ability to form hydrogen bonds. | Pharmacophore descriptors | (+) ve | |
| Neoplastic−50 | This descriptor indicates the Ghose-Viswanadhan-Wendoloski antineoplastic-like index at 50 %. It is generated by analysing the dataset of antineoplastic drugs, where 50 % indicates 50 % of the drugs follow a particular range of physicochemical properties. Any chemical that falls within this range may also be likely to have similar antineoplastic potential. The ranges of the physicochemical properties for the Neoplastic−50 descriptor are: ALOGP, 0.0–3.7; AMR, 60–107; MW, 258–388; and nAT, 30–55. [69] | Drug-like indices | (+) ve | |

(*continued on next page*)

**Table 2** (*continued*)

| Descriptor | Explanation | Class | Direction of contribution | Target Tissue |
|---|---|---|---|---|
| | The preceding discussion indicates that these compounds exhibit lower lipophilicity and larger molecular sizes. Consequently, such compounds possess an enhanced potential to partition into peripheral tissues. Notably, muscle tissue is characterized by a lower fat content and a higher water content. [72,74] As a result, these compounds demonstrate a greater propensity to partition within muscle tissue, which is indicated by their positive contributions. | | | |
| NsssN | This descriptor indicates the number of atom types of the single bond, single bond, and single bond nitrogen atom. The presence of these electronegative atoms renders the molecule hydrophilic. Since skin is rich in lipids, such fragments in a molecule decrease its ability to partition into the skin. [75] | Atom-type E-state indices | (-) ve | **Skin** |
| Mor16i | It refers to the signal or index number 16 in the MoRSE descriptor series, where i indicates that the descriptor is weighted by ionization potential. | 3D-MoRSE descriptors | (+) ve | |
| B08[C-N] | This descriptor indicates the presence or absence of carbon and nitrogen atoms at a topological distance of eight. It indicates the presence of large fragments within a molecule that increase its lipophilicity, which in turn enhances the partitioning of the drug. | 2D Atom Pairs | (+) ve | **Spleen** |

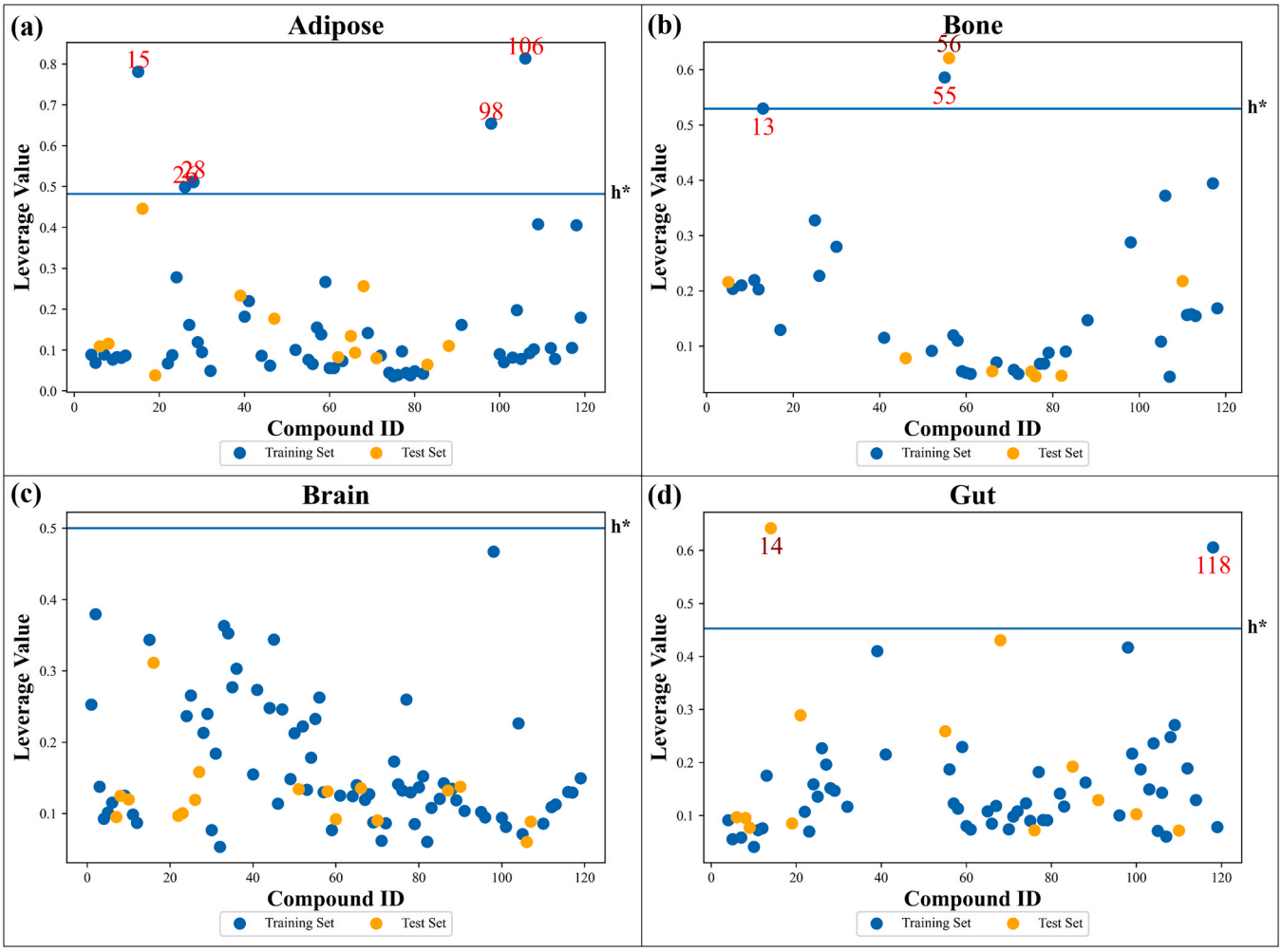



**Fig. 5.** Applicability domain (AD) plot. (a) AD plot for adipose. (b) AD plot for bone. (c) AD plot for the brain. (d) AD plot for gut.

physicochemical properties of the drugs, which demonstrated significantly high quality ($R^2$ = 0.83–0.95) and predictivity ($Q^2_{F1}$ = 0.71–0.93 and $Q^2_{F2}$ = 0.70–0.92). These models were further refined by incorporating $K_p$ values from other tissues, as a few organs share similar physiological conditions and are influenced by each other. From these models, we found that predictivity ($Q^2_{F1}$ = 0.79–0.95 and $Q^2_{F2}$ = 0.78–0.95) improved significantly. However, the quality ($R^2$ = 0.85–0.96) of the models remained almost the same, with a similar or lower number of features.

In this study, we conducted a screening using a true external set obtained from the SIDER database. The data we collected includes eight different categories of toxicity information. We applied our final selected models to predict the tissue-to-plasma partition coefficient of a particular tissue for these compounds. The results indicate that compounds with a greater tendency to partition in a specific tissue also tend to exhibit toxicity within that same tissue.

Finally, to enhance user convenience, we developed a Python-based graphical user interface for predicting the $K_p$ parameter of drugs. This tool estimates the $K_p$ parameter along with the applicability domain status by simply giving the input of the descriptor matrix. The only

**Table 3**
Top four compounds that show higher partitioning for a particular tissue and also exhibit toxicity for that tissue (SIDER database).

| Organ Toxicity | Top Four Compounds |
|---|---|
| Skin and Subcutaneous Tissue Disorders | Atracurium, Mepenzolate, Umeclidinium bromide, Fingolimod hydrochloride |
| Renal and Urinary Disorders | Umeclidinium bromide, CHEMBL1201344, Mepenzolate, Ipratropium |
| Musculoskeletal or Tissue Disorder | BDBM86425, HMS1361O16, CHEMBL1201344, 9-chloro−5-(4-hydroxyphenyl)−2,3,4,5-tetrahydro-1H−3-benzazepin−3-ium−7,8-diol |
| Hepatobiliary Disorders | HMS1361O16, Methylthioninium, Erythromid, Ioxilan |
| Cardiac Disorders | Atracurium besylate, Umeclidinium bromide, Doxorubicin, Mepenzolate |
| Nervous System Disorders | Cyclobenzaprine, HMS1361O16, Atovaquone, Perhexiline |
| Gastrointestinal Disorders | C6V6LNU2YL, Lurasidone, Deferoxamine, Dabigatran etexilate |
| Respiratory Disorders | CHEMBL1201344, HMS1361O16, Cetylpyridinium, BDBM86425 |

limitation of this tool is that the descriptors must be computed using a different software.

This study provides a reliable alternative approach for predicting the $K_p$ of a new chemical in a fast and cost-effective manner. These findings not only help to avoid unnecessary animal testing but also reduce the time involved in animal experimentation. The developed models could be further improved by incorporating new data in the future.

## CRediT authorship contribution statement

**Kunal Roy:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization. **Souvik Pore:** Writing – original draft, Validation, Software, Investigation, Formal analysis, Data curation.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Table 4**
Comparative analysis between previous studies and current studies.

| Tissue | Previous studies | | Current studies |
|---|---|---|---|
| | Study 1 [36] | Study 2 [37] | |
| **Adipose** | $R^2 = 0.67$<br>RMSE = 0.47 | $R^2 = 0.47$<br>RMSE = 0.49 | $R^2 = 0.89$<br>RMSEC = 0.33<br>$Q^2_{F1} = 0.79$<br>RMSEP = 0.39 |
| **Bone** | $R^2 = 0.36$<br>RMSE = 0.51 | - | $R^2 = 0.87$<br>RMSEC = 0.36<br>$Q^2_{F1} = 0.85$<br>RMSEP = 0.43 |
| **Brain** | $R^2 = 0.63$<br>RMSE = 0.46 | $R^2 = 0.52$<br>RMSE = 0.50 | $R^2 = 0.89$<br>RMSEC = 0.32<br>$Q^2_{F1} = 0.82$<br>RMSEP = 0.44 |
| **Gut** | $R^2 = 0.70$<br>RMSE = 0.38 | $R^2 = 0.76$<br>RMSE = 0.37 | $R^2 = 0.91$<br>RMSEC = 0.29<br>$Q^2_{F1} = 0.95$<br>RMSEP = 0.22 |
| **Heart** | $R^2 = 0.68$<br>RMSE = 0.33 | $R^2 = 0.64$<br>RMSE = 0.34 | $R^2 = 0.92$<br>RMSEC = 0.28<br>$Q^2_{F1} = 0.95$<br>RMSEP = 0.21 |
| **Kidney** | $R^2 = 0.68$<br>RMSE = 0.34 | $R^2 = 0.75$<br>RMSE = 0.34 | $R^2 = 0.87$<br>RMSEC = 0.35<br>$Q^2_{F1} = 0.91$<br>RMSEP = 0.39 |
| **Liver** | $R^2 = 0.52$<br>RMSE = 0.48 | $R^2 = 0.52$<br>RMSE = 0.46 | $R^2 = 0.85$<br>RMSEC = 0.38<br>$Q^2_{F1} = 0.90$<br>RMSEP = 0.31 |
| **Lungs** | $R^2 = 0.84$<br>RMSE = 0.36 | $R^2 = 0.75$<br>RMSE = 0.42 | $R^2 = 0.95$<br>RMSEC = 0.23<br>$Q^2_{F1} = 0.94$<br>RMSEP = 0.27 |
| **Muscle** | $R^2 = 0.68$<br>RMSE = 0.32 | $R^2 = 0.70$<br>RMSE = 0.34 | $R^2 = 0.88$<br>RMSEC = 0.35<br>$Q^2_{F1} = 0.91$<br>RMSEP = 0.32 |
| **Skin** | $R^2 = 0.64$<br>RMSE = 0.28 | $R^2 = 0.51$<br>RMSE = 0.29 | $R^2 = 0.90$<br>RMSEC = 0.31<br>$Q^2_{F1} = 0.94$<br>RMSEP = 0.25 |
| **Spleen** | $R^2 = 0.73$<br>RMSE = 0.39 | - | $R^2 = 0.96$<br>RMSEC = 0.19<br>$Q^2_{F1} = 0.89$<br>RMSEP = 0.26 |

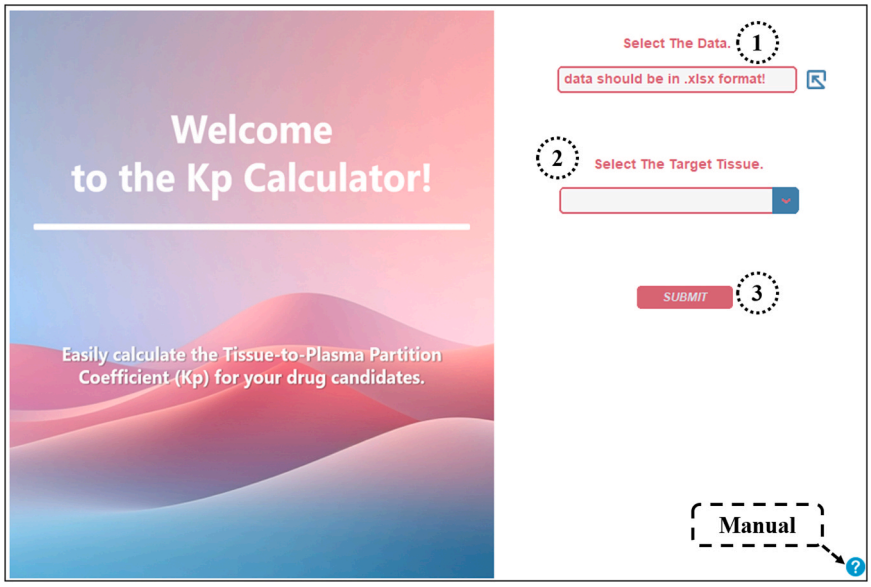



**Fig. 6.** Graphical user interface.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.aichem.2025.100093.

## Data availability statement

The structural information of the molecules used for the modelling is available in Supplementary Material 3 (SM3). The machine learning modelling tools used in this study are available from the official website of the DTC Lab tool (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis). The input files for machine learning model development are available in Supplementary Material 2 (SM2). The developed prediction tool is available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/kp-calculator.

## References


[1] N. Berdigaliyev, M. Aljofan, An overview of drug discovery and development, Future Med. Chem. 12 (10) (2020) 939–947, https://doi.org/10.4155/FMC-2019-0307.

[2] J. Vrbanac, R. SlauterADME in Drug Discovery. A Comprehensive Guide to Toxicology in Nonclinical Drug Development 2017, 39–67. https://doi.org/10.1016/B978-0-12-803620-4.00003-7.2017.

[3] G.C. Wallace, D.B. Ramsden, M.H. Grant, General principles of drug distribution, Encycl. Drug Metab. Interact. (2012) 1–47, https://doi.org/10.1002/9780470921920.EDM024.

[4] D.K. Walker, The use of pharmacokinetic and pharmacodynamic data in the assessment of drug safety in early drug development, Br. J. Clin. Pharmacol. 58 (6) (2004) 601–608, https://doi.org/10.1111/J.1365-2125.2004.02194.X.

[5] M.J. Waring, J. Arrowsmith, A.R. Leach, P.D. Leeson, S. Mandrell, R.M. Owen, G. Pairaudeau, W.D. Pennie, S.D. Pickett, J. Wang, O. Wallace, A. Weir, An analysis of the attrition of drug candidates from four major pharmaceutical companies, Nat. Rev. Drug Discov. 14 (7) (2015) 475–486, https://doi.org/10.1038/nrd4609.

[6] J.M. Gallo, Pharmacokinetic/ pharmacodynamic-driven drug development, Mt. Sinai J. Med. A J. Transl. Pers. Med. 77 (4) (2010) 381–388, https://doi.org/10.1002/MSJ.20193.

[7] S.Q. Pantaleão, P.O. Fernandes, J.E. Gonçalves, V.G. Maltarollo, K.M. Honorio, Recent advances in the prediction of pharmacokinetics properties in drug design studies: a review, ChemMedChem 17 (1) (2022) e202100542, https://doi.org/10.1002/CMDC.202100542.

[8] A. Gruber, F. Führer, S. Menz, H. Diedam, A.H. Göller, S. Schneckener, Prediction of human pharmacokinetics from chemical structure: combining mechanistic modeling with machine learning, J. Pharm. Sci. 113 (1) (2024) 55–63, https://doi.org/10.1016/J.XPHS.2023.10.035.

[9] X. Zhuang, C. Lu, PBPK modeling and simulation in drug research and development, Acta Pharm. Sin. B 6 (5) (2016) 430–440, https://doi.org/10.1016/J.APSB.2016.04.004.

[10] D. Deepika, V. Kumar, The role of "Physiologically based pharmacokinetic model (PBPK)" new approach methodology (NAM) in pharmaceuticals and environmental chemical risk assessment, Int. J. Environ. Res. Public Health 20 (4) (2023) 3473, https://doi.org/10.3390/IJERPH20043473.

[11] H.M. Jones, K. Rowland-Yeo, Basic concepts in physiologically based pharmacokinetic modeling in drug discovery and development, CPT Pharmacomet. Syst. Pharmacol. 2 (8) (2013) 1–12, https://doi.org/10.1038/PSP.2013.41.

[12] H. Graham, M. Walker, O. Jones, J. Yates, A. Galetin, L. Aarons, Comparison of In-Vivo and in-Silico methods used for prediction of tissue: plasma partition coefficients in rat, J. Pharm. Pharmacol. 64 (3) (2012) 383–396, https://doi.org/10.1111/J.2042-7158.2011.01429.X.

[13] P. Poulin, F.P. Theil, A priori prediction of Tissue:Plasma partition coefficients of drugs to facilitate the use of Physiologically-Based pharmacokinetic models in drug discovery, J. Pharm. Sci. 89 (1) (2000) 16–35, https://doi.org/10.1002/(sici)1520-6017(200001)89:1<16::aid-jps3>3.0.co;2-e.

[14] W. Schmitt, General approach for the calculation of tissue to plasma partition coefficients, Toxicol. Vitr. 22 (2) (2008) 457–467, https://doi.org/10.1016/J.TIV.2007.09.010.

[15] J.H. Lin, Y. Sugiyama, S. Awazu, M. Hanano, In vitro and in vivo evaluation of the Tissue-to-Blood partition coefficient for physiological pharmacokinetic models, J. Pharm. Biopharm. 10 (6) (1982) 637–647, https://doi.org/10.1007/BF01062545/METRICS.

[16] Y.E. Yun, C.A. Cotton, A.N. Edginton, Development of a decision tree to classify the most accurate Tissue-Specific tissue to plasma partition coefficient algorithm for a given compound, J. Pharm. Pharm. 41 (1) (2014) 1–14, https://doi.org/10.1007/S10928-013-9342-0/FIGURES/3.

[17] H. Graham, M. Walker, O. Jones, J. Yates, A. Galetin, L. Aarons, Comparison of In-Vivo and in-Silico methods used for prediction of tissue: plasma partition coefficients in rat, J. Pharm. Pharmacol. 64 (3) (2012) 383–396, https://doi.org/10.1111/J.2042-7158.2011.01429.X.

[18] R.W. Johnson, N. Talaty, Tissue imaging by mass spectrometry: a practical guide for the medicinal chemist, ACS Med. Chem. Lett. 10 (2) (2019) 161–167, https://doi.org/10.1021/ACSMEDCHEMLETT.8B00480/ASSET/IMAGES/LARGE/ML-2018-00480F_0001.JPEG.

[19] P. Poulin, F.P. Theil, A priori prediction of Tissue:Plasma partition coefficients of drugs to facilitate the use of Physiologically-Based pharmacokinetic models in drug discovery, J. Pharm. Sci. 89 (1) (2000) 16–35, https://doi.org/10.1002/(sici)1520-6017(200001)89:1<16::aid-jps3>3.0.co;2-e.

[20] Poulin, P.; Schoenlein, K.; Theil, F.-P.; Hoffmann-La, F. Prediction of Adipose Tissue:Plasma Partition Coef®cients for Structurally Unrelated Drugs. https://doi.org/10.1002/1520-6017.

[21] S. Björkman, Reduction and lumping of physiologically based pharmacokinetic models: prediction of the disposition of fentanyl and pethidine in humans by successively simplified models, J. Pharm. Pharm. 30 (4) (2003) 285–307, https://doi.org/10.1023/A:1026194618660/METRICS.

[22] S. Pore, K. Roy, Insights into pharmacokinetic properties for exposure chemicals: predictive modelling of human plasma fraction unbound ( f u) and hepatocyte intrinsic clearance (Cl Int) data using machine learning, Digit. Discov. 3 (9) (2024) 1852–1877, https://doi.org/10.1039/D4DD00082J.

[23] C. Ogami, Y. Tsuji, H. Seki, H. Kawano, H. To, Y. Matsumoto, H. Hosono, An artificial neural network−pharmacokinetic model and its interpretation using shapley additive explanations, CPT Pharmacomet. Syst. Pharmacol. 10 (7) (2021) 760–768, https://doi.org/10.1002/PSP4.12643.

[24] O. Obrezanova, Artificial intelligence for compound pharmacokinetics prediction, Curr. Opin. Struct. Biol. 79 (2023) 102546, https://doi.org/10.1016/J.SBI.2023.102546.

[25] F. Miljković, A. Martinsson, O. Obrezanova, B. Williamson, M. Johnson, A. Sykes, A. Bender, N. Greene, Machine learning models for human in vivo pharmacokinetic parameters with In-House validation, Mol. Pharm. 18 (12) (2021) 4520–4530, https://doi.org/10.1021/ACS.MOLPHARMACEUT.1C00718/ASSET/IMAGES/LARGE/MP1C00718_0007.JPEG.

[26] Y. Wang, H. Liu, Y. Fan, X. Chen, Y. Yang, L. Zhu, J. Zhao, Y. Chen, Y. Zhang, In silico prediction of human intravenous pharmacokinetic parameters with improved accuracy, J. Chem. Inf. Model 59 (9) (2019) 3968–3980, https://doi.org/10.1021/ACS.JCIM.9B00300/ASSET/IMAGES/LARGE/CI9B00300_0009.JPEG.

[27] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao, Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov. 18 (6) (2019) 463–477, https://doi.org/10.1038/s41573-019-0024-5.

[28] S. Dara, S. Dhamercherla, S.S. Jadav, C.M. Babu, M.J. Ahsan, Machine learning in drug discovery: a review, Artif. Intell. Rev. 55 (3) (2021) 1947–1999, https://doi.org/10.1007/S10462-021-10058-4.

[29] L. Patel, T. Shukla, X. Huang, D.W. Ussery, S. Wang, Machine learning methods in drug discovery, Molecules 25 (22) (2020) 5277, https://doi.org/10.3390/MOLECULES25225277.

[30] H. Gummesson Svensson, C. Tyrchan, O. Engkvist, M. Haghir Chehreghani, Utilizing reinforcement learning for de novo drug design, Mach. Learn 113 (7) (2024) 4811–4843, https://doi.org/10.1007/S10994-024-06519-W/FIGURES/22.

[31] G. Obaido, I.D. Mienye, O.F. Egbelowo, I.D. Emmanuel, A. Ogunleye, B. Ogbuokiri, P. Mienye, K. Aruleba, Supervised machine learning in drug discovery and development: algorithms, applications, challenges, and prospects, Mach. Learn. Appl. 17 (2024) 100576, https://doi.org/10.1016/J.MLWA.2024.100576.

[32] A. Géron, Hands-on machine learning with Scikit-Learn, keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, O'Reilly Media Inc. (2022).

[33] J. Chen, H. Yang, L. Zhu, Z. Wu, W. Li, Y. Tang, G. Liu, In silico prediction of human renal clearance of compounds using quantitative Structure-Pharmacokinetic relationship models, Chem. Res. Toxicol. 33 (2) (2020) 640–650, https://doi.org/10.1021/ACS.CHEMRESTOX.9B00447/ASSET/IMAGES/LARGE/TX9B00447_0008.JPEG.

[34] K. Roy, S. Kar, R.N. Das, Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment, Academic press, 2015.

[35] K. Mayumi, M. Tachibana, M. Yoshida, S. Ohnishi, T. Kanazu, H. Hasegawa, The novel in vitro method to calculate Tissue-to-Plasma partition coefficient in humans for predicting pharmacokinetic profiles by Physiologically-Based pharmacokinetic model with high predictability, J. Pharm. Sci. 109 (7) (2020) 2345–2355, https://doi.org/10.1016/J.XPHS.2020.04.002.

[36] K. Handa, S. Sakamoto, M. Kageyama, T. Iijima, Development of a 2D-QSAR model for Tissue-to-Plasma partition coefficient value with high accuracy using machine learning method, minimum required experimental values, and physicochemical descriptors, Eur. J. Drug Metab. Pharm. 48 (4) (2023) 341–352, https://doi.org/10.1007/S13318-023-00832-W/FIGURES/4.

[37] K. Handa, S. Yoshimura, M. Kageyama, T. Iijima, Development of novel methods for QSAR modeling by machine learning repeatedly: a case study on drug distribution to each tissue, J. Chem. Inf. Model 64 (9) (2024) 3662–3669, https://doi.org/10.1021/ACS.JCIM.4C00046/ASSET/IMAGES/LARGE/CI4C00046_0004.JPEG.

[38] Y.E. Yun, A.N. Edginton, Correlation-based prediction of tissue-to-plasma partition coefficients using readily available input parameters, Xenobiotica 43 (10) (2013) 839–852, https://doi.org/10.3109/00498254.2013.770182.

[39] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open babel: an open chemical toolbox, J. Cheminf. 3 (10) (2011) 1–14, https://doi.org/10.1186/1758-2946-3-33/TABLES/2.

[40] V. Consonni, R. Todeschini, Molecular descriptors, Chall. Adv. Comput. Chem. Phys. 8 (2010) 29–102, https://doi.org/10.1007/978-1-4020-9783-6_3.

[41] P. Gramatica, On the development and validation of QSAR models, Methods Mol. Biol. 930 (2013) 499–526, https://doi.org/10.1007/978-1-62703-059-5_21.
[42] A. Mauri, AlvaDesc: a tool to calculate and analyze molecular descriptors and fingerprints, Methods Pharmacol. Toxicol. (2020) 801–820, https://doi.org/10.1007/978-1-0716-0150-1_32.
[43] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inf. 29 (6–7) (2010) 476–488, https://doi.org/10.1002/MINF.201000061.
[44] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model 52 (10) (2012) 2570–2578, https://doi.org/10.1021/CI300338W/SUPPL_FILE/CI300338W_SI_001.PDF.
[45] P.P. Roy, J.T. Leonard, K. Roy, Exploring the impact of size of training sets for the development of predictive QSAR models, Chemom. Intell. Lab. Syst. 90 (1) (2008) 31–42, https://doi.org/10.1016/J.CHEMOLAB.2007.07.004.
[46] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1) (1969) 137–148, https://doi.org/10.1080/00401706.1969.10490666.
[47] P. Ambure, R.B. Aher, A. Gajewicz, T. Puzyn, K. Roy, NanoBRIDGES" software: open access tools to perform QSAR and Nano-QSAR modeling, Chemom. Intell. Lab. Syst. 147 (2015) 1–13, https://doi.org/10.1016/J.CHEMOLAB.2015.07.007.
[48] I.S. Oh, J.S. Lee, B.R. Moon, Hybrid genetic algorithms for feature selection, IEEE Trans. Pattern Anal. Mach. Intell. 26 (11) (2004) 1424–1437, https://doi.org/10.1109/TPAMI.2004.105.
[49] D. Rogers, A.J. Hopfinger, Application of genetic function approximation to quantitative structure-activity relationships and quantitative Structure-Property relationships, J. Chem. Inf. Comput. Sci. 34 (4) (1994) 854–866, https://doi.org/10.1021/CI00020A020/ASSET/CI00020A020.FP.PNG_V03.
[50] L. Breiman, Random forests, Mach. Learn 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324/METRICS.
[51] Y. Freund, R.E. Schapire, A short introduction to boosting, J. Jpn. Soc. Artif. Intell. 14 (5) (1999) 771–780.
[52] Freund, Y.; Schapire, R.E. Experiments with a New Boosting Algorithm. Proc 13th Int Conf Mach Learn 1996. https://doi.org/https://doi.org/10.1.1.133.1040.
[53] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, 13-17-August-2016, 785–794. https://doi.org/10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.
[54] S. Suthaharan, Support Vector Machine, 36, Springer Nature Link, 2016, pp. 207–235, https://doi.org/10.1007/978-1-4899-7641-3_9.
[55] G.C. McDonald, Ridge regression, Wiley Inter. Rev. Comput. Stat. 1 (1) (2009) 93–100, https://doi.org/10.1002/WICS.14.
[56] V. Esposito Vinzi, G. Russolillo, Partial least squares algorithms and methods, Wiley Inter. Rev. Comput. Stat. 5 (1) (2013) 1–19, https://doi.org/10.1002/WICS.1239.
[57] I.H. Sarker, Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions, SN Comput. Sci. 2 (6) (2021) 1–20, https://doi.org/10.1007/S42979-021-00815-1/FIGURES/11.
[58] L.M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R.M. Blair, W.S. Branham, S.L. Dial, C.L. Moland, D.M. Sheehan, QSAR models using a large diverse set of estrogens, J. Chem. Inf. Comput. Sci. 41 (1) (2001) 186–195, https://doi.org/10.1021/CI000066D/ASSET/IMAGES/LARGE/CI000066DF00005.JPEG.
[59] G. Schüürmann, R.U. Ebert, J. Chen, B. Wang, R. Kühne, External validation and prediction employing the predictive squared correlation coefficient - test set activity mean vs training set activity mean, J. Chem. Inf. Model 48 (11) (2008) 2140–2145, https://doi.org/10.1021/CI800253U/ASSET/IMAGES/CI-2008-00253U_M013.GIF.
[60] Lundberg, S.M.; Allen, P.G.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions.
[61] L. Shapley, 7. A value for n-Person games. Contributions to the theory of games II, Class. Game Theory (1953) 69–79, https://doi.org/10.1515/9781400829156-012/HTML.
[62] R. Rodríguez-Pérez, J. Bajorath, Interpretation of machine learning models using shapley values: application to compound potency and Multi-Target activity predictions, J. Comput. Aided Mol. Des. 34 (10) (2020) 1013–1026, https://doi.org/10.1007/S10822-020-00314-0/FIGURES/8.
[63] S. Pore, A. Banerjee, K. Roy, Machine Learning-Based q-RASPR modeling of power conversion efficiency of organic dyes in Dye-Sensitized solar cells, Sustain. Energy Fuels 7 (14) (2023) 3412–3431, https://doi.org/10.1039/D3SE00457K.
[64] S. Pore, A. Pelloux, M. Chatterjee, A. Banerjee, K. Roy, Machine Learning-Based q-RASAR predictions of the bioconcentration factor of organic molecules estimated following the organisation for economic Co-Operation and development guideline 305, J. Hazard. Mater. 479 (2024) 135725, https://doi.org/10.1016/J.JHAZMAT.2024.135725.
[65] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of different approaches to define the applicability domain of QSAR models, Molecules 17 (5) (2012) 4791–4810, https://doi.org/10.3390/MOLECULES17054791.
[66] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemom. Intell. Lab. Syst. 145 (2015) 22–29, https://doi.org/10.1016/J.CHEMOLAB.2015.04.013.
[67] N. Minovski, Š. Župerl, V. Drgan, M. Novič, Assessment of applicability domain for multivariate Counter-Propagation artificial neural network predictive models by minimum Euclidean distance space analysis: a case study, Anal. Chim. Acta 759 (2013) 28–42, https://doi.org/10.1016/J.ACA.2012.11.002.
[68] http://sideeffects.embl.de/ (Accessed on 31 July 2025).
[69] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, A Knowledge-Based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. a qualitative and quantitative characterization of known drug databases, J. Comb. Chem. 1 (1) (1999) 55–68, https://doi.org/10.1021/CC9800071/ASSET/IMAGES/LARGE/CC9800071F00005.JPEG.
[70] X. Hou, L. Zhang, Z. Zhou, X. Luo, T. Wang, X. Zhao, B. Lu, F. Chen, L. Zheng, Calcium phosphate-based biomaterials for bone repair, J. Funct. Biomater. 13 (4) (2022) 187, https://doi.org/10.3390/JFB13040187.
[71] A.L. Boskey, P.G. Robey, The composition of bone, Prim. Metab. Bone Dis. Disord. Miner. Metab. Eighth Ed. (2013) 49–58, https://doi.org/10.1002/9781118453926.CH6.
[72] T.F. Tirrell, M.S. Cook, J.A. Carr, E. Lin, S.R. Ward, R.L. Lieber, Human skeletal muscle biochemical diversity, J. Exp. Biol. 215 (15) (2012) 2551, https://doi.org/10.1242/JEB.069385.
[73] F.A. Duck, Tissue composition, Phys. Prop. Tissues (1990) 319–328, https://doi.org/10.1016/B978-0-12-222800-1.50013-9.
[74] I. Lorenzo, M. Serra-Prat, Carlos Yébenes, J. The role of water homeostasis in muscle function and frailty: a review, Nutrients 11 (8) (2019) 1857, https://doi.org/10.3390/NU11081857.
[75] D.J. Tobin, Biochemistry of human Skin—our brain on the outside, Chem. Soc. Rev. 35 (1) (2006) 52–67.
[76] R. Parker, The role of adipose tissue in fatty liver diseases, Liver Res. 2 (1) (2018) 35–42, https://doi.org/10.1016/J.LIVRES.2018.02.002.
[77] I. Tattoli, A.R. Mathew, A. Verrienti, L. Pallotta, C. Severi, F. Andreola, V. Cavallucci, M. Giorgi, M. Massimi, L. Bencini, M. Fidaleo, The interplay between liver and adipose tissue in the onset of liver diseases: exploring the role of vitamin deficiency, Cells 13 (19) (2024) 1631, https://doi.org/10.3390/CELLS13191631.
[78] M.S. Roberts, B.M. Magnusson, F.J. Burczynski, M. Weiss, Enterohepatic circulation, Clin. Pharmacokinet. 41 (10) (2012) 751–790, https://doi.org/10.2165/00003088-200241100-00005.
[79] J.M. Padowski, G.M. Pollack, Influence of enterohepatic recycling on the time course of Brain-to-Blood partitioning of valproic acid in rats, Drug Metab. Dispos. 40 (9) (2012) 1846–1853, https://doi.org/10.1124/DMD.112.045500.
[80] J. Guan, C. Wu, Y. He, F. Lu, Skin-Associated adipocytes in skin barrier immunity: a mini-review, Front. Immunol. 14 (2023) 1116548, https://doi.org/10.3389/FIMMU.2023.1116548/BIBTEX.
[81] R. Darlenski, V. Mihaylova, T. Handjieva-Darlenska, The link between obesity and the skin, Front. Nutr. 9 (2022) 855573, https://doi.org/10.3389/FNUT.2022.855573/BIBTEX.