**Skin permeability prediction for molecules using various AI algorithms: a comparative study for different molecular representations**

Rami M. Abdallah\*1, Hisham E. Hasan1, Jehad M. Nasereddin\*1, Ahmad Hammad2

1- Department of Pharmaceutical Sciences, Faculty of Pharmacy, Zarqa University, Zarqa, Jordan

2- Department of Artificial Intelligence, Faculty of Information Technology, Middle East University, Amman, Jordan

\* **Corresponding Author**

Jehad M. Nasereddin, PhD; Assistant Professor of Pharmaceutical Sciences, Faculty of Pharmacy, Zarqa University, P.O. Box 13024, Amman. 11942, Jordan. Tel: +962 79 850 0900, work: +9625-3821100, extensions 1584, Email: [jnasereddin@zu.edu.jo](mailto:jnasereddin@zu.edu.jo)

**Conflict of Interest**

The authors declare that there is no conflict of interest.

**Authors’ Contribution**

All the authors contributed equally to this paper.

**Data Availability Statement**

The raw data supporting the conclusions of this article will be made available by the authors.

**Supplementary Material**

The Supplementary Material for this article can be found online at: --------------

**Funding Sources**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for profit organisations.

**Acknowledgments**

We would like to thank the following volunteers

**Abstract**

In recent years, the skin has emerged as a convenient route of drug administration for local and systemic therapeutic effects, but effective drug delivery through the skin is challenging due to the skin itself acting as a natural barrier. Empirical rules for selecting drugs for transdermal drug delivery (TDD) are limited and provide only qualitative estimations of permeability, making it difficult to compare between drugs. Therefore, there is a need for appropriate models that can provide accurate quantitative predictions for the skin permeability of drugs. Here, we present the development of a novel regression model using multiple artificial intelligence (AI) algorithms and distinct molecular representations, such as descriptor-based and graph-based models. Additionally, we aim to investigate the permeability patterns of FDA-approved drugs to advance our current understanding of drug permeability.

**Introduction**

Recently, skin has started to receive more interest as a convenient route of drug administration for both local and systemic therapeutic effects, thus it has been frequently selected for the development of drug delivery systems (Alkilani et al., 2022; Jeong et al., 2021). However, several challenges are present for the effective delivery of drugs through the skin, which forms a natural barrier for the permeation of xenobiotics, and the development of complicated pharmaceutical technology (i.e., transdermal drug delivery, which is of interest) is even more challenging (Alkilani et al., 2015). The most notable one of them is that the drug must have suitable physicochemical properties to enable it to penetrate the stratum corneum and reach the bloodstream with a sufficient dose (Souto et al., 2022). Currently, formulation scientists are investigating the right drug for transdermal drug delivery (TDD) by relying on empirical rules where plenty of drugs show deviations from them and show varied behavior (Phatale et al., 2022; Yu et al., 2021). Moreover, these rules provide a qualitative estimation of permeability, and it is generally difficult to compare drugs that obey these rules. Therefore, with the accumulated data on the permeability of drugs, it is essential to develop appropriate models that provide accurate quantitative predictions for the skin permeability of drugs (Baba et al., 2015).

*In silico* QSPR (Quantitative Structure-Property Relationship), which can help predict how novel compounds will behave, uses numerical descriptors for the structure of molecules to correlate them with a specific property (Wu et al., 2022). QSPR was extensively used to develop predictive models for skin permeability, commonly using linear regression (LR) and principal component analysis (PCA) models (Geinoz et al., 2004; Tsakovska et al., 2017). However, due to the complex and diverse nature of the chemicals involved, nonlinear regression methods such as support vector machine (SVM), random forest (RF), and artificial neural networks (ANN) are more appealing solutions than other linear methods like LR and PCA, as they are often more effective at identifying patterns and nonlinear relationships within complex datasets (Baba et al., 2015; Wu et al., 2022).

Artificial intelligence (AI) is a rapidly evolving field of research that seeks to design and build machines with the ability to complete tasks that typically require human intelligence, such as problem-solving, learning, and decision-making. The potential benefits of AI are vast and can include improving healthcare and education, among many other areas (Xu et al., 2021). In the context of drug delivery, AI can be used to develop models that can predict the permeability and bioavailability of drugs based on their physicochemical properties. This has the potential to greatly accelerate the drug development process and help identify promising drug candidates that may have otherwise been overlooked (Paul et al., 2021). Several AI models have been developed to estimate skin permeability from basic physicochemical characteristics (Agatonovic-Kustrin et al., 2020; Atobe et al., 2015; Baba et al., 2015; Bušatlić et al., 2017; Chen et al., 2007; Deği̇m et al., 2003; Lim et al., 2002). The majority of them rely on calculated descriptors for model training in order to determine the permeability coefficient (kp).

ANN uses data modeling and pattern recognition to simulate the learning and generalization behavior of the human brain for complex multidimensional problems. They are commonly employed for the purpose of modeling and predicting the properties and behavior of molecules (Walters & Barzilay, 2021). However, a significant limitation of this approach lies in the fact that ANNs are typically trained on descriptors, which do not always provide a complete representation of the underlying molecular structure. This can lead to the loss of important information and potential inaccuracies in the predictions. To address this issue, a promising alternative is to use the molecular structure itself as input to the AI model. Graph neural networks (GNNs) are a subtype of ANNs that are well-suited to handling molecule structures represented as graphs. (Zhou et al., 2020). By directly incorporating the graph structure into the model, GNNs are able to capture more detailed information about the molecules, which can lead to more accurate predictions (Jiang et al., 2021). This is demonstrated by several studies that show GNN outperforming or at least equally performing descriptors-based models. However, performance comparisons between GNN and descriptor-based models are still a topic of debate and may need evaluation on a case-by-case basis.

The objective of this study is to develop a regression model employing diverse AI algorithms, including GNN, that can predict kp for new compounds solely based on their structure. To the best of our knowledge, this is the first instance of applying the GNN algorithm to predict the skin permeability of molecules. The model will also be used for predicting kp of FDA-approved drugs, and cluster analysis will be conducted to classify drugs based on the selected features and kp into classes.

**Methods**

***Dataset***

In this study, a skin permeability dataset was acquired from the work of Cheruvu et al (Cheruvu et al., 2022) comprising a diverse range of molecules such as drugs, xenobiotics, and other chemical compounds. The dataset provides the logKp values for the tested molecules against human epidermal membranes. The SMILES structures of the compounds were retrieved from PubChem for generating descriptors in subsequent analyses.

***Calculation of Descriptors***

In this study, the generation of descriptors for the molecules was performed using the open-source, Java-based chemoinformatics library Chemistry Development Kit (CDK) version 2.8 (Willighagen et al., 2017). The SMILES structure of the molecules was imported into the library, and a comprehensive set of 1D/2D descriptors was calculated for subsequent use as inputs in the AI models. Prior to descriptor calculation, water was excluded, and salts were neutralized. To handle missing data, the columns containing errors were filled with the mean value of that respective column.

***AI Models***

In that study, two types of experiments were conducted, namely regression analysis and cluster analysis. The experiments were carried out using Scikit-Learn version 1.2 (Pedregosa et al., 2011), which was an open-source library utilizing the Python programming language. The process involved various stages, including data cleaning, preprocessing, and splitting into training and testing sets. Afterward, the model was trained, a baseline model was selected, feature selection was performed, and hyperparameters were tuned. These steps were repeated iteratively.

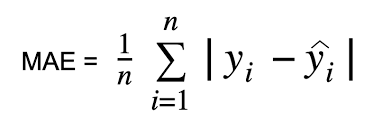
***Regression Analysis***

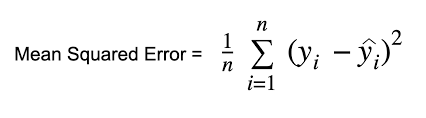
We used regression analysis, which is a supervised learning approach, to predict kp for molecules. We tried various machine learning models, including multiple linear regression, support vector machines, and ensemble algorithms, whether it was a bagging or boosting algorithm, such as random forest, XGBoost, CatBoost, and LGBM, as well as using ANN and GNN.

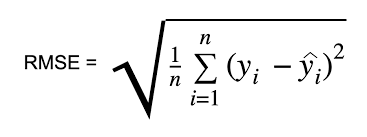
***Cluster Analysis***

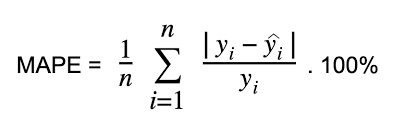
The permeability of FDA-approved drugs was predicted, and cluster analysis, which is an unsupervised learning approach and data mining technique that can find similarities between data points, was conducted. Fortunately, we already had the target variable, which provided a clear direction for the study. We used K-means Clustering.

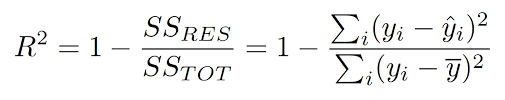
***Metrics***

******

******

******



******

<https://www.google.com/imgres?imgurl=https%3A%2F%2Feditor.analyticsvidhya.com%2Fuploads%2F42439Screenshot%25202021-10-26%2520at%25209.34.08%2520PM.png&tbnid=0HpzJ_8fF3BKMM&vet=12ahUKEwjNvNXOrvL-AhV2rycCHYIHC1sQMygAegQIARBP..i&imgrefurl=https%3A%2F%2Fwww.analyticsvidhya.com%2Fblog%2F2021%2F10%2Fevaluation-metric-for-regression-models%2F&docid=6Swd6ZI0TQ0q_M&w=722&h=250&q=MAE%20formula%20in%20paper&ved=2ahUKEwjNvNXOrvL-AhV2rycCHYIHC1sQMygAegQIARBP>

**Results**

***Characterization of The Dataset***

The dataset consists of 476 records for 145 different molecules with a diverse range of LogKp values, spanning from -5.53 cm/h to -0.08 cm/h, and was generated under varying temperatures, ranging from 295 K to 312 K. The distribution of LogKp values, molecular weight, LogP, water solubility and melting point are depicted in Figures 1-5, which provide a comprehensive characterization of the dataset. Specifically, Figure 1 illustrates the distribution of LogKp values, while Figures 2, 3, and 4 exhibit the distribution of molecular weight, LogP, water solubility, and melting point, respectively. The dataset exhibits a substantial coverage of a diverse range of molecules, indicating its suitability for conducting rigorous analyses and development for regression models.

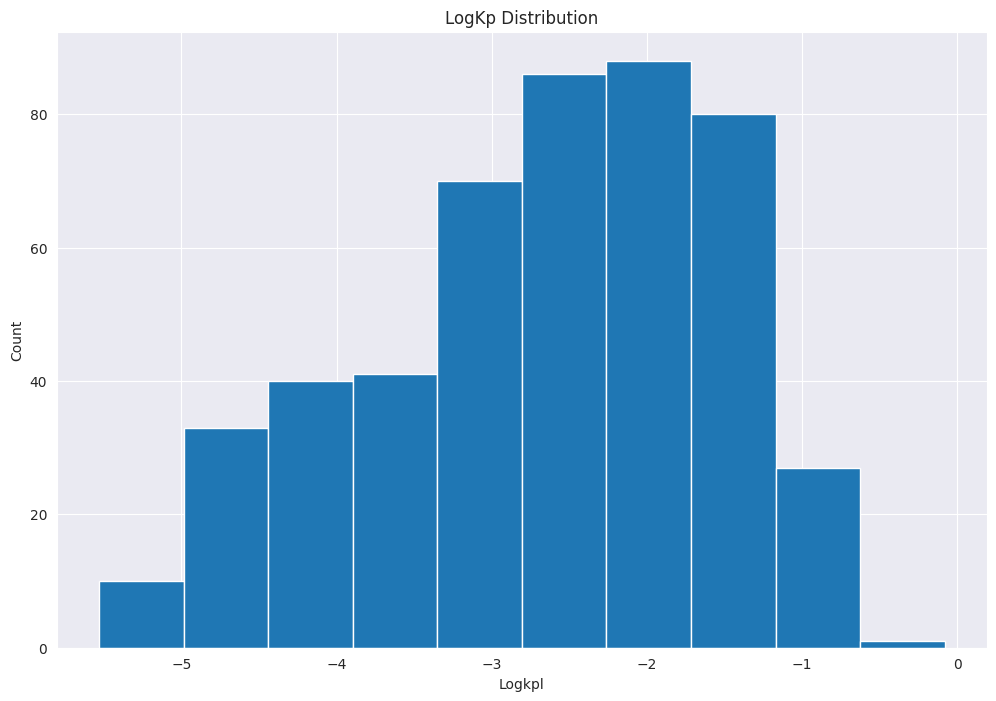


Figure 1 LogKp distribution for the dataset

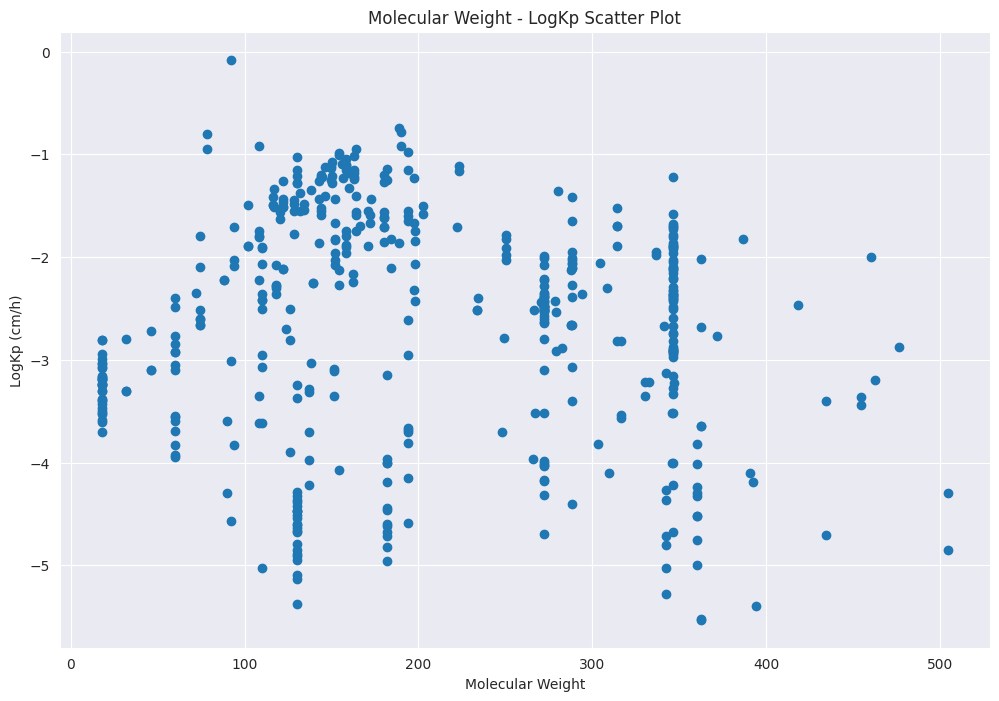


Figure 2

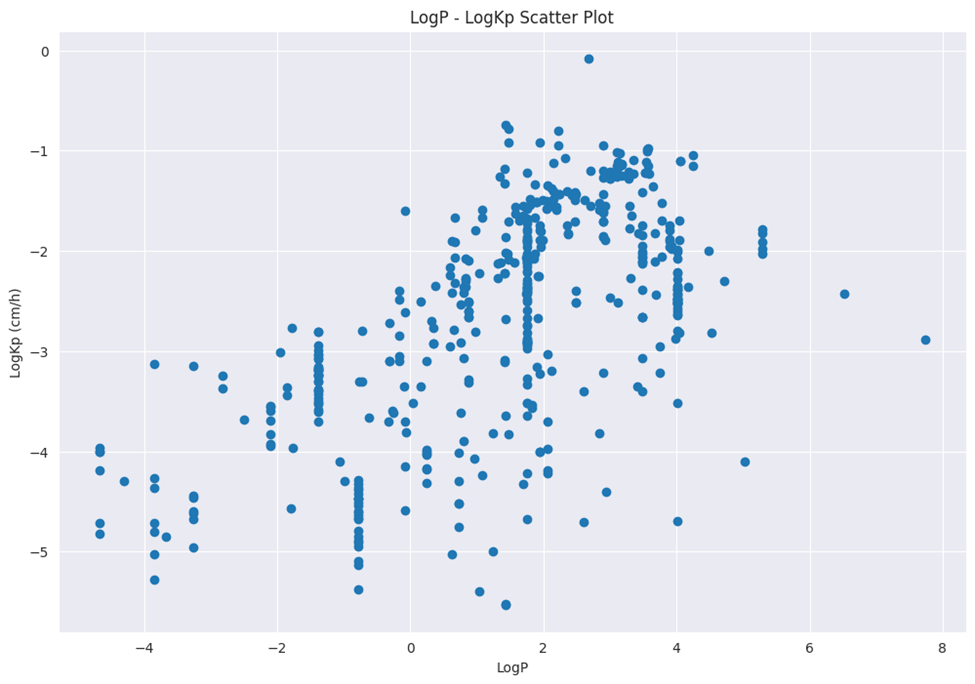


Figure 3

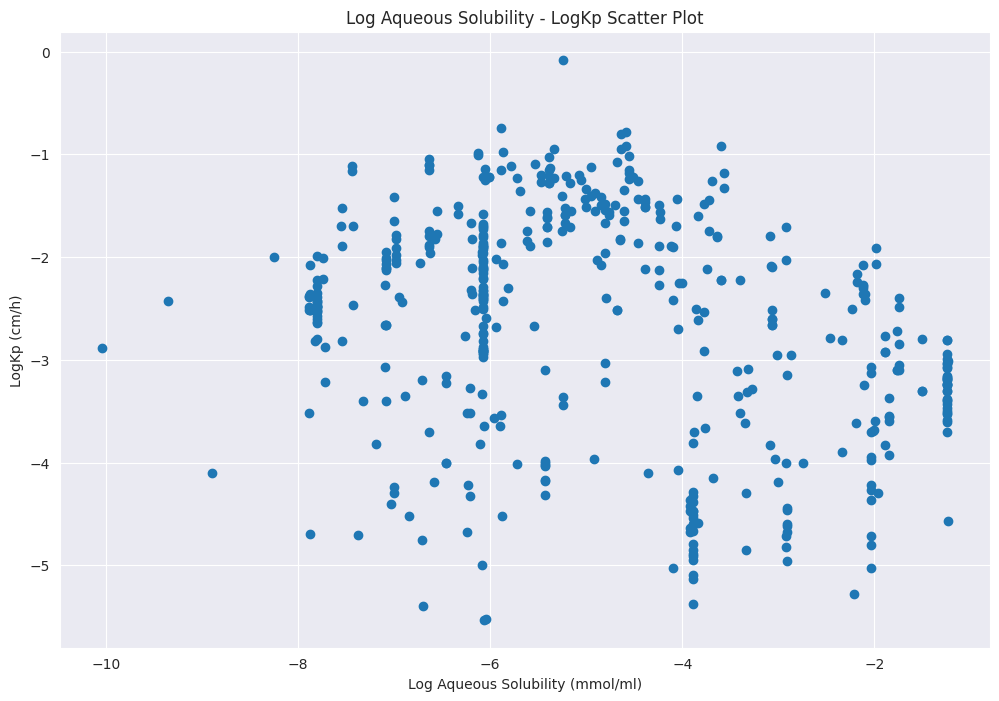


Figure 4

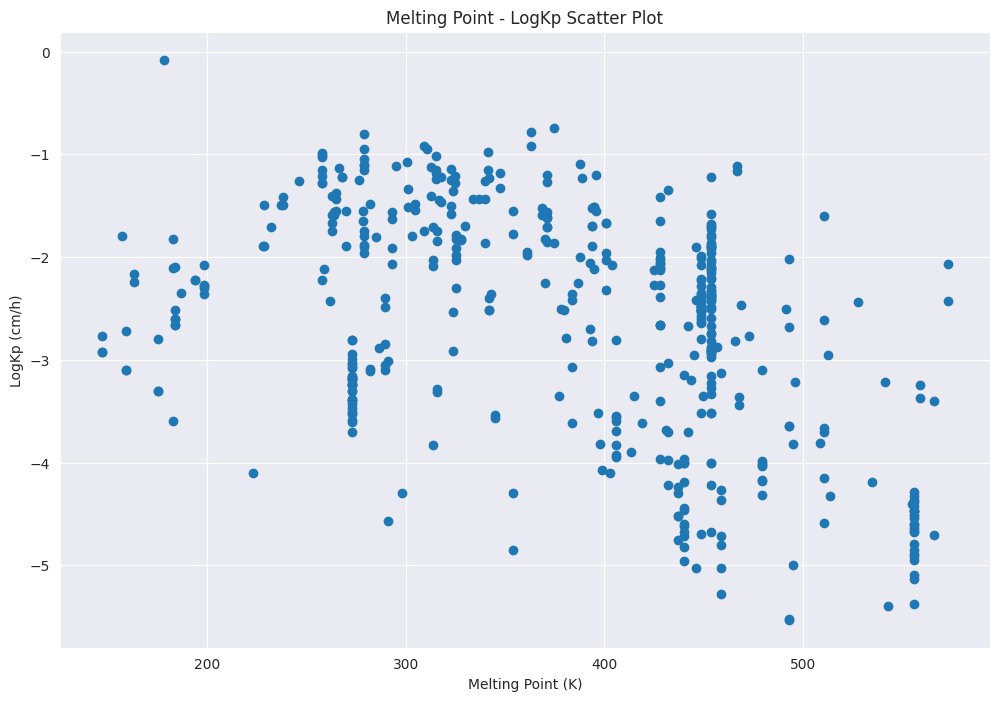


Figure 5

***Descriptors Calculation***

A total of 222 1D/2D descriptors were computed for the compounds present in the dataset. The descriptors heatmap (Figure 6) provides an overview of the correlations between the different molecular descriptors used in this study. As expected, there are strong correlations between certain descriptors, such as molecular weight and polar surface area (PSA), as well as between LogP and the number of rotatable bonds. These correlations suggest that molecules with higher molecular weight tend to have larger PSA values, while those with higher LogP tend to have a higher number of rotatable bonds. Additionally, there are weaker correlations between other pairs of descriptors, such as between LogP and molecular weight, and between PSA and the number of hydrogen bond acceptors.

Interestingly, the heatmap also reveals some unexpected correlations, such as a moderate positive correlation between LogP and the number of hydrogen bond donors. This suggests that molecules with higher LogP may also have more hydrogen bond donors, which could impact their binding properties and solubility. Overall, the heatmap provides valuable insights into the relationships between different molecular descriptors, which can help guide the development of predictive models for drug design and discovery.

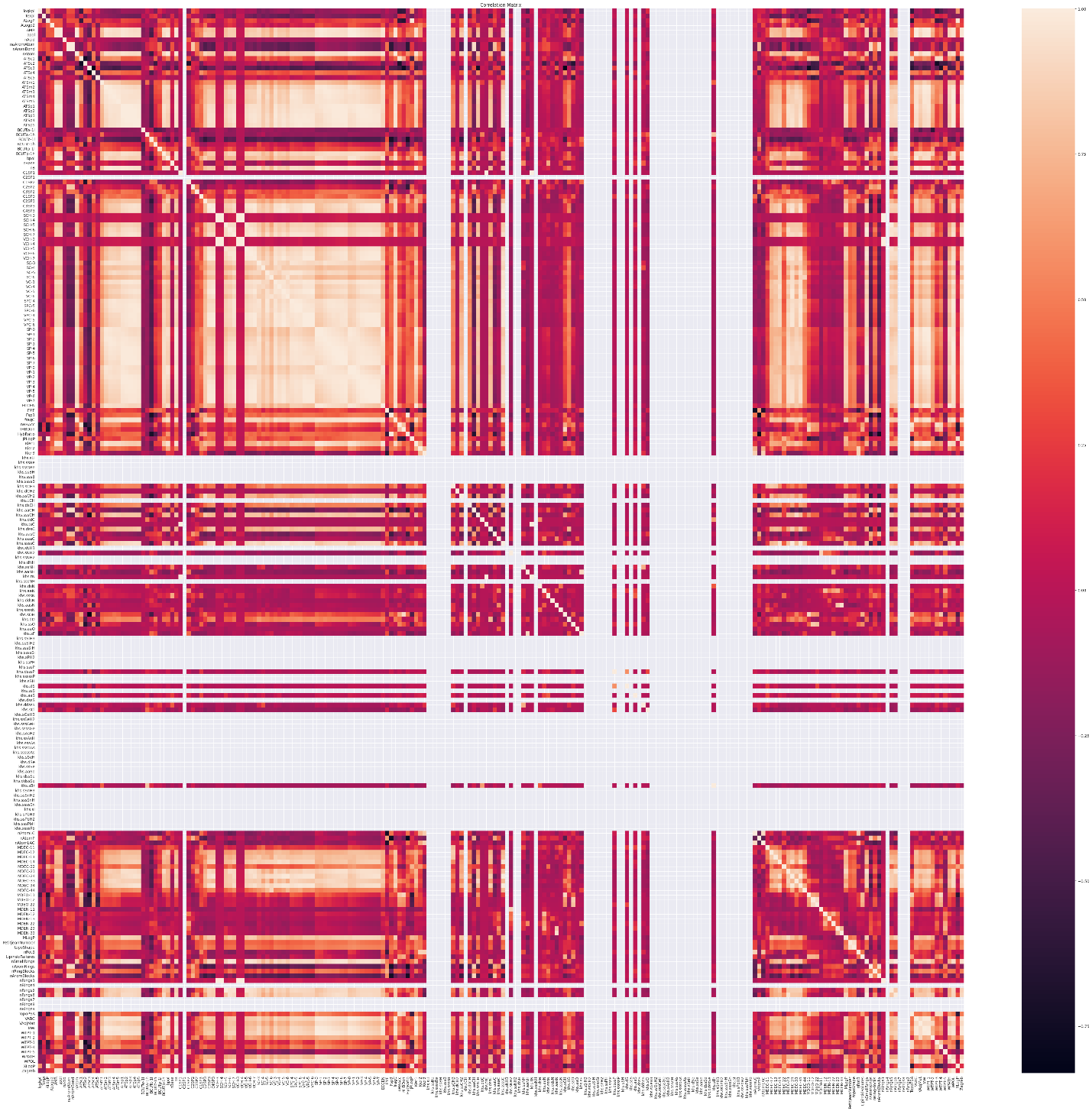


Figure 6

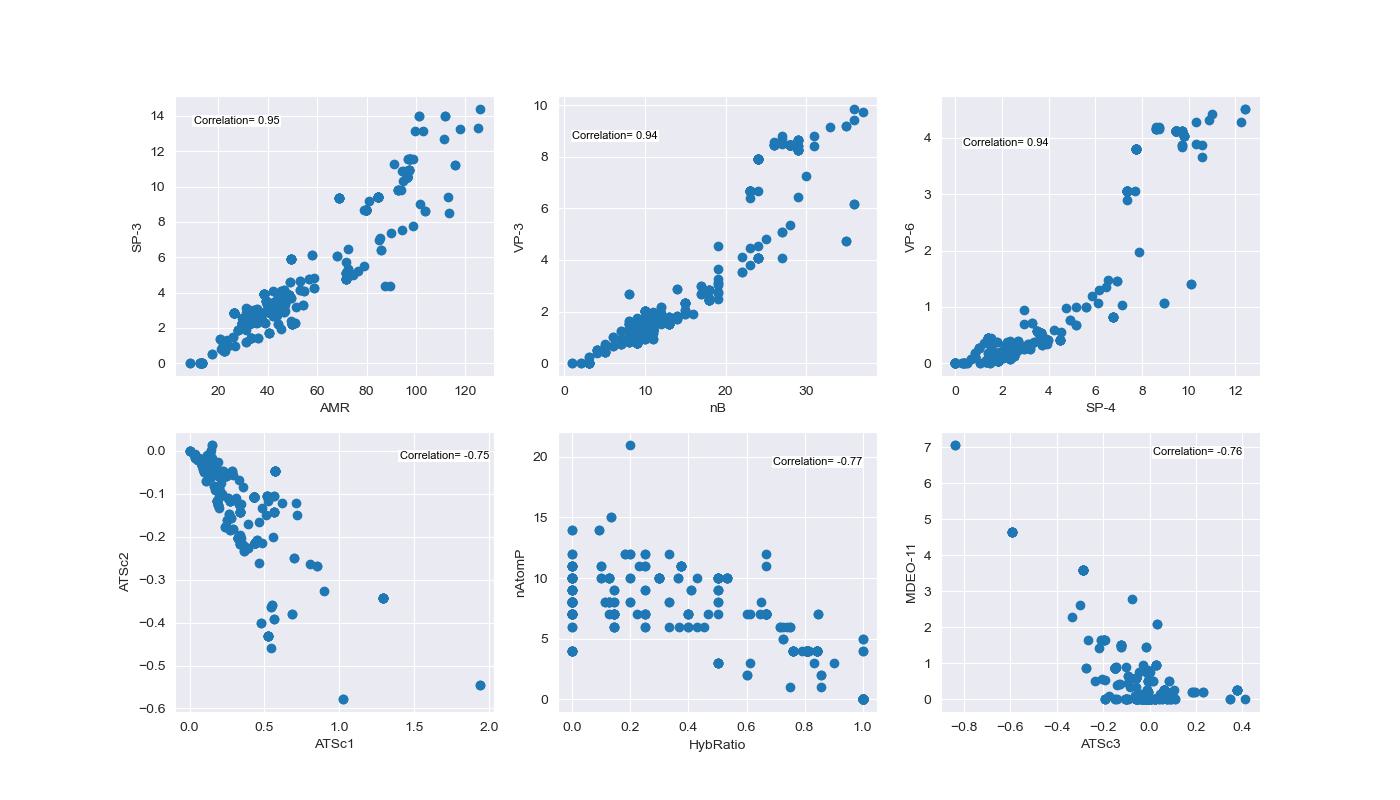


Figure 7: Some Correlated Features

Figure 7 illustrates some correlated features with, we can see that SP-3 is highly positively correlated with AMR, same for VP-3 with nB, and VP-6 with SP-4, there are highly negatively correlated features such as ATSc2 with ATSc1, nAtomP with HybRatio, and MDEO-11 with ATSC3.

***Regression Models Development***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **MSE** | **MAE** | **MAPE** | **RMSE** | **R2** |
| Linear Regression | 15.85 | 1.11 | 0.56 | 3.98 | -12.89 |
| Random Forest | 0.29 | 0.39 | 0.18 | 0.54 | 0.74 |
| **XGBoost** | **0.26** | **0.35** | **0.16** | **0.51** | **0.77** |
| Gradient Boosting | 0.29 | 0.39 | 0.18 | 0.54 | 0.75 |
| CatBoost | 0.28 | 0.37 | 0.18 | 0.53 | 0.76 |
| LGBM | 0.29 | 0.38 | 0.18 | 0.54 | 0.75 |
| ANN | 0.48 | 0.48 | 0.24 | 0.69 | 0.58 |
| GNN | 0.55 | 0.66 | 0.34 | 0.74 | 0.07 |

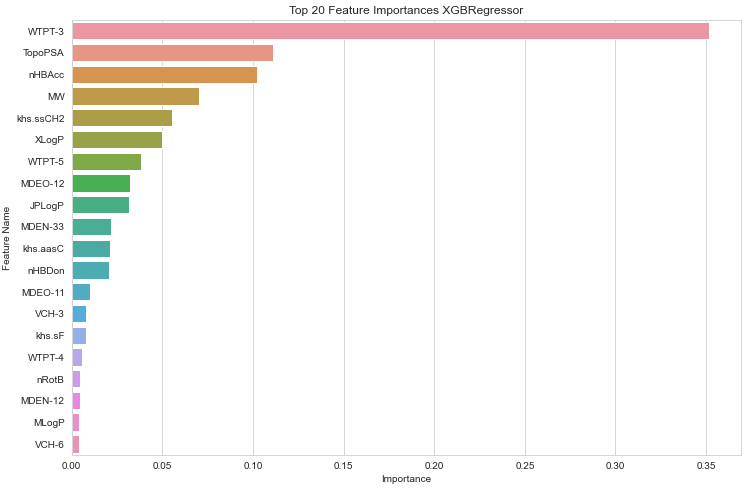
The regression analysis was performed using various machine learning models, including multiple linear regression, random forest, XGBoost, CatBoost, LGBM, ANN, and GNN. Table 2 shows the performance of the models based on their root mean squared error (RMSE), mean absolute error (MAE), and R2 score.

Based on the metrics reported in the table, it appears that the Random Forest, XGBoost, CatBoost, and Gradient Boosting models performed better than the other models in terms of MSE, MAE, MAPE, and RMSE. Among these four models, XGBoost had the lowest MSE, MAE, MAPE, and RMSE, and the highest R2 value, indicating that it performed the best overall. The ANN model had the highest MSE, MAE, MAPE, and RMSE, and the lowest R2 value, indicating that it performed the worst among the models.

The GNN model also had relatively high MSE, MAE, MAPE, and RMSE values, and a low R2 value, suggesting that it may not be the best choice for this particular task.

Based on the performance metrics obtained for the different models, it can be concluded that the Random Forest, XGBoost, Gradient Boosting, CatBoost, and LGBM models outperformed the Linear Regression, ANN, and GNN models in predicting the target variable.

The XGBoost model achieved the lowest MSE, RMSE, MAE and MAPE values. The XGBoost and CatBoost models achieved the highest R2 values, indicating that they were able to explain a larger portion of the variance in the target variable.

   
ANN Architecture  
Decision Tree Outline  
Representation of different number of neurons in hidden layer vs performance  
test set performance scatter plot for best model  
top features image  
RMS functions written

***Cluster Analysis***



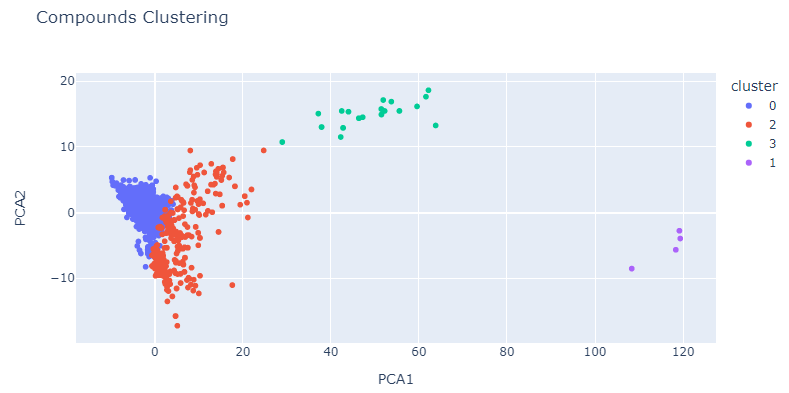
*Figure 8: Cluster Analysis with compounds*

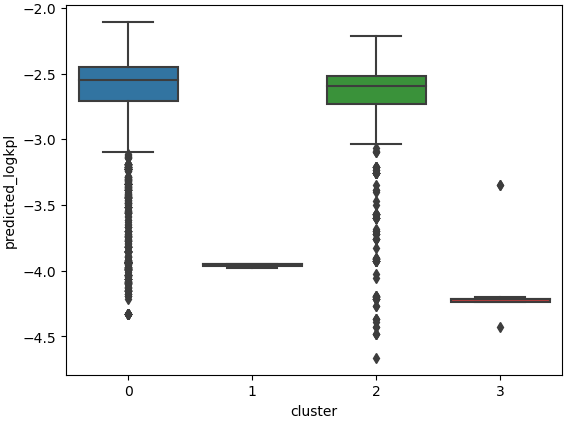
A picture containing diagram, rectangle, square, plan

Description automatically generated

*Figure 9: LogKp for Each Cluster*

We used 3k drugs approved data.

*Figure 10: Cluster Analysis with compounds*



*Figure 11: Cluster Analysis with compounds*

**Discussion**

The present study investigated the

**Conclusion**

This study provides valuable insights

**References**

Agatonovic-Kustrin, S., Chan, C. K. Y., Gegechkori, V., & Morton, D. W. (2020). Models for skin and brain penetration of major components from essential oils used in aromatherapy for dementia patients. *Journal of Biomolecular Structure and Dynamics*, *38*(8), 2402–2411. https://doi.org/10.1080/07391102.2019.1633408

Alkilani, A. Z., McCrudden, M. T. C., & Donnelly, R. F. (2015). Transdermal drug delivery: Innovative pharmaceutical developments based on disruption of the barrier properties of the stratum corneum. In *Pharmaceutics* (Vol. 7, Issue 4, pp. 438–470). MDPI AG. https://doi.org/10.3390/pharmaceutics7040438

Alkilani, A. Z., Nasereddin, J., Hamed, R., Nimrawi, S., Hussein, G., Abo-Zour, H., & Donnelly, R. F. (2022). Beneath the Skin: A Review of Current Trends and Future Prospects of Transdermal Drug Delivery Systems. In *Pharmaceutics* (Vol. 14, Issue 6). MDPI. https://doi.org/10.3390/pharmaceutics14061152

Atobe, T., Mori, M., Yamashita, F., Hashida, M., & Kouzuki, H. (2015). Artificial neural network analysis for predicting human percutaneous absorption taking account of vehicle properties. *The Journal of Toxicological Sciences*, *40*(2), 277–294. https://doi.org/10.2131/jts.40.277

Baba, H., Takahara, J. I., & Mamitsuka, H. (2015). In silico predictions of human skin permeability using nonlinear quantitative structure-property relationship models. *Pharmaceutical Research*, *32*(7), 2360–2371. https://doi.org/10.1007/s11095-015-1629-y

Bušatlić, E., Osmanović, A., Jakupović, A., Nuhić, J., & Hodžić, A. (2017). Using neural networks and ensemble techniques based on decision trees for skin permeability prediction. *IFMBE Proceedings*, *62*, 41–50. https://doi.org/10.1007/978-981-10-4166-2\_8

Chen, L. J., Lian, G. P., & Han, L. J. (2007). Prediction of human skin permeability using artificial neural network (ANN) modeling. *Acta Pharmacologica Sinica*, *28*(4), 591–600. https://doi.org/10.1111/j.1745-7254.2007.00528.x

Cheruvu, H. S., Liu, X., Grice, J. E., & Roberts, M. S. (2022). An updated database of human maximum skin fluxes and epidermal permeability coefficients for drugs, xenobiotics, and other solutes applied as aqueous solutions. *Data in Brief*, *42*, 108242. https://doi.org/10.1016/J.DIB.2022.108242

Deği̇m, T., Hadgraft, J., İlbasmiş, S., & Özkan, Y. (2003). Prediction of skin penetration using artificial neural network (ANN) modeling. *Journal of Pharmaceutical Sciences*, *92*(3), 656–664. https://doi.org/https://doi.org/10.1002/jps.10312

Geinoz, S., Guy, R. H., Testa, B., & Carrupt, P.-A. (2004). *Quantitative Structure-Permeation Relationships (QSPeRs) to Predict Skin Permeation: A Critical Evaluation*.

Jeong, W. Y., Kwon, M., Choi, H. E., & Kim, K. S. (2021). Recent advances in transdermal drug delivery systems: a review. In *Biomaterials Research* (Vol. 25, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s40824-021-00226-6

Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, *13*(1). https://doi.org/10.1186/s13321-020-00479-8

Lim, C. W., Fujiwara, S., Yamashita, F., & Hashida, M. (2002). Prediction of human skin permeability using a combination of molecular orbital calculations and artificial neural network. *Biological and Pharmaceutical Bulletin*, *25*(3), 361–366.

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. In *Drug Discovery Today* (Vol. 26, Issue 1, pp. 80–93). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2020.10.010

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Phatale, V., Vaiphei, K. K., Jha, S., Patil, D., Agrawal, M., & Alexander, A. (2022). Overcoming skin barriers through advanced transdermal drug delivery approaches. In *Journal of Controlled Release* (Vol. 351, pp. 361–380). Elsevier B.V. https://doi.org/10.1016/j.jconrel.2022.09.025

Souto, E. B., Fangueiro, J. F., Fernandes, A. R., Cano, A., Sanchez-Lopez, E., Garcia, M. L., Severino, P., Paganelli, M. O., Chaud, M. v., & Silva, A. M. (2022). Physicochemical and biopharmaceutical aspects influencing skin permeation and role of SLN and NLC for skin drug delivery. In *Heliyon* (Vol. 8, Issue 2). Elsevier Ltd. https://doi.org/10.1016/j.heliyon.2022.e08938

Tsakovska, I., Pajeva, I., al Sharif, M., Alov, P., Fioravanzo, E., Kovarich, S., Worth, A. P., Richarz, A. N., Yang, C., Mostrag-Szlichtyng, A., & Cronin, M. T. D. (2017). Quantitative structure-skin permeability relationships. In *Toxicology* (Vol. 387, pp. 27–42). Elsevier Ireland Ltd. https://doi.org/10.1016/j.tox.2017.06.008

Walters, W. P., & Barzilay, R. (2021). Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Accounts of Chemical Research*, *54*(2), 263–270. https://doi.org/10.1021/acs.accounts.0c00699

Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C. T., Guha, R., & Steinbeck, C. (2017). The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics*, *9*(1). https://doi.org/10.1186/s13321-017-0220-4

Wu, Y. W., Ta, G. H., Lung, Y. C., Weng, C. F., & Leong, M. K. (2022). In Silico Prediction of Skin Permeability Using a Two-QSAR Approach. *Pharmaceutics*, *14*(5). https://doi.org/10.3390/pharmaceutics14050961

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C. W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., … Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. In *The Innovation* (Vol. 2, Issue 4). Cell Press. https://doi.org/10.1016/j.xinn.2021.100179

Yu, Y. Q., Yang, X., Wu, X. F., & Fan, Y. bin. (2021). Enhancing Permeation of Drug Molecules Across the Skin via Delivery in Nanocarriers: Novel Strategies for Effective Transdermal Applications. In *Frontiers in Bioengineering and Biotechnology* (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fbioe.2021.646554

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. In *AI Open* (Vol. 1, pp. 57–81). Elsevier B.V. https://doi.org/10.1016/j.aiopen.2021.01.001