**Dataset PK I: Regression models for prediction of human intravenous pharmacokinetic parameters**

Size: A dataset of 1352 compounds with human intravenous pharmacokinetic parameters

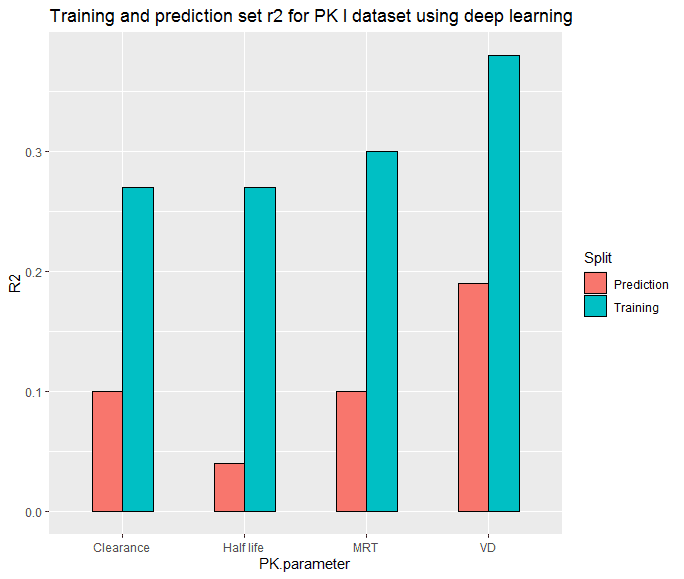
Enumerated parameters: Terminal half life(t1/2,), renal clearance (Clren), mean residence time(MRT), steady state volume of distribution (Vdss)

Reference: Dataset adapted from Obach *et al*., 2018. (doi: [10.1124/dmd.118.082966](https://dmd.aspetjournals.org/content/suppl/2018/08/16/dmd.118.082966.DC1))

Modeling approach: Deep learning (Graph neural networks) and multiple linear regression

1. **Deep learning using Graph Neural Networks**

**Table 1:** Validation metrics for PKI (deep learning models)



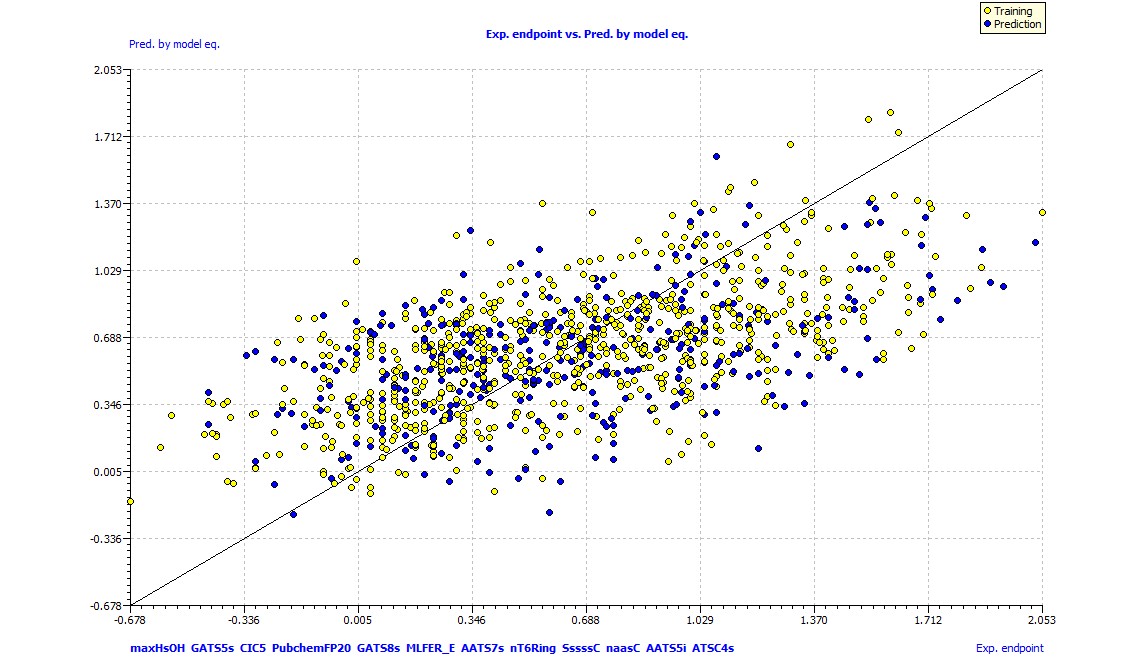
**Figure 1:** Comparison of training and test r2 for PK I (deep learning models**)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training/Test | R2 | R2test |
| Half life | 839/359 | 0.27 | 0.04 |
| Clearance | 839/359 | 0.27 | 0.1 |
| MRT | 839/359 | 0.3 | 0.1 |
| VD | 839/359 | 0.38 | 0.19 |

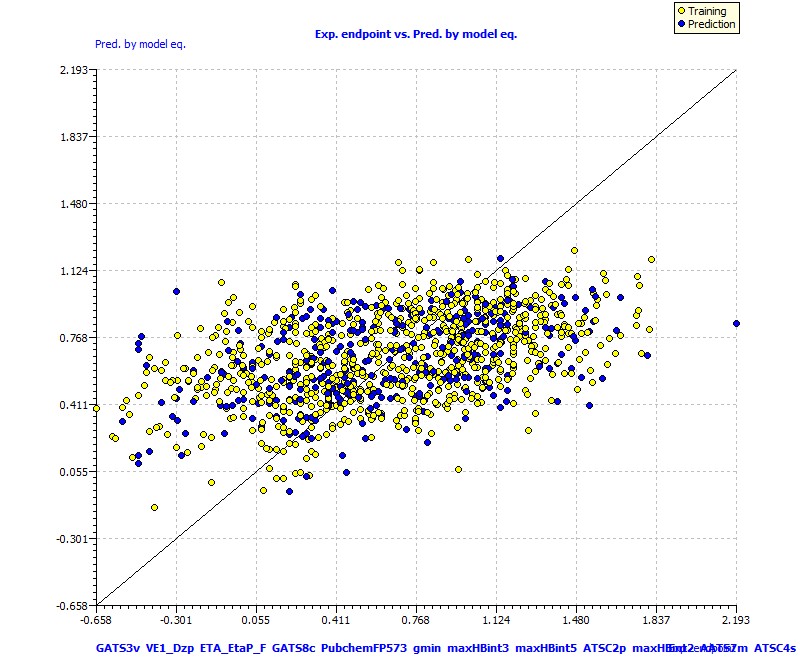
1. **Multiple linear regression**

**Table 2:** Computed validation metrics for PK I (MLR models)

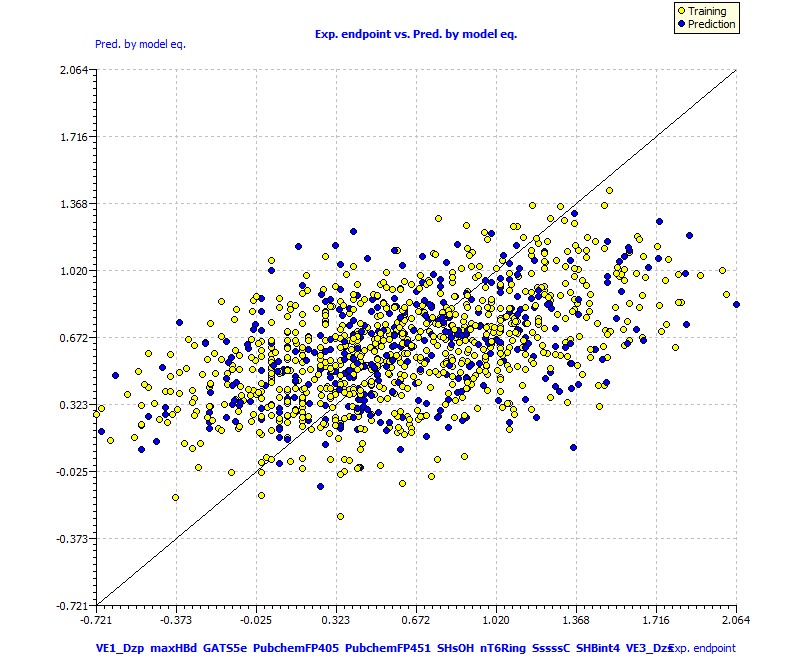
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Training/Test | R2 | R2adj | RMSE tr | MAE tr | Q2loo |  | R2 Yscr | RMSE ext | MAE ext | R2ext | r2m aver. |
| Half life | 777/330 | 0.3895 | 0.3799 | 0.3955 | 0.321 | 0.3664 |  | 0.0156 | 0.4334 | 0.3493 | 0.2468 | 0.1278 |
| Clearance | 794/328 | 0.2155 | 0.2034 | 0.4429 | 0.3605 | 0.1901 |  | 0.0149 | 0.4436 | 0.3566 | 0.1901 | 0.0525 |
| MRT | 786/337 | 0.2777 | 0.2683 | 0.4452 | 0.3619 | 0.2591 |  | 0.0129 | 0.45 | 0.3599 | 0.2148 | 0.0928 |
| VD | 776/331 | 0.589 | 0.5826 | 0.3674 | 0.2948 | 0.5744 |  | 0.0156 | 0.4098 | 0.3311 | 0.4907 | 0.3284 |



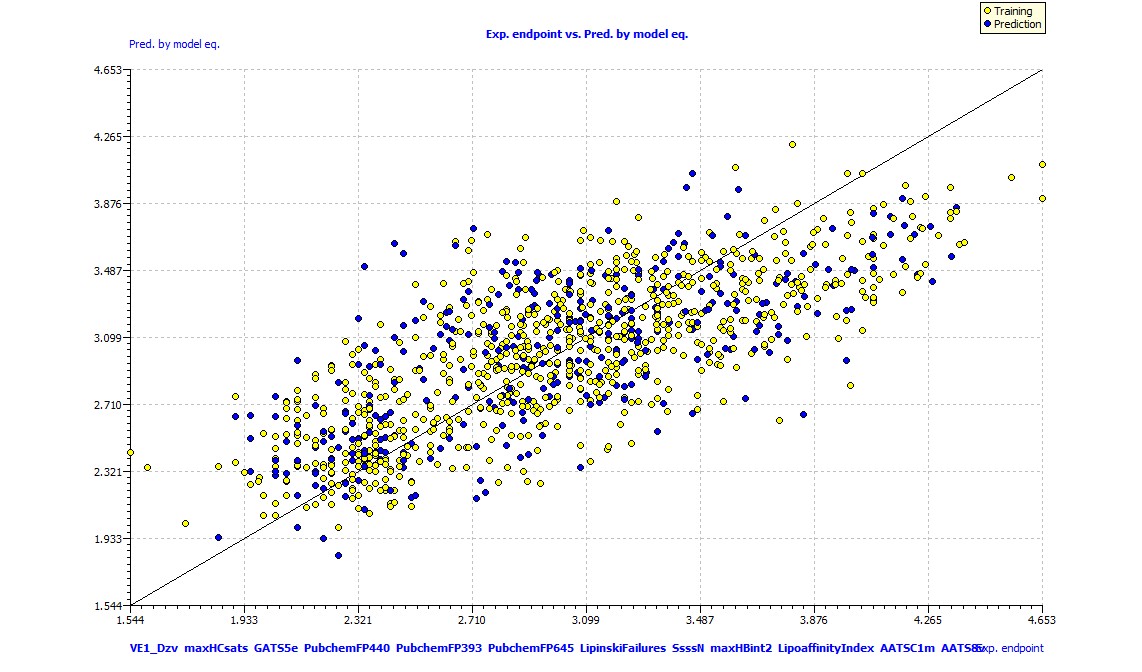
**Figure 2:** Exp endpoint vs Pred by model plot for half-life (MLR model)



**Figure 3:** Exp endpoint vs Pred by model plot for clearance (MLR model)



**Figure 4:** Exp endpoint vs Pred by model plot for MRT (MLR model)



**Figure 5:** Exp endpoint vs Pred by model plot for Vd (MLR model)

**Dataset PK II- Classification models for Blood-Brain Barrier Penetration**

Size: 605 compounds with experimental BBB permeability

Reference: Dataset adapted from Singh *et al*., 2020 (doi: [10.1016/j.jmgm.2019.107516](https://doi.org/10.1016/j.jmgm.2019.107516))

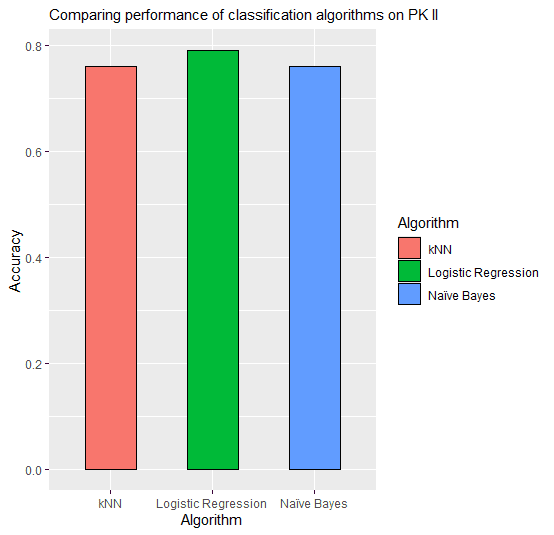
Modeling approach: Binary classification using machine learning algorithms (Logistic regression, naïve bayes classifier, k nearest neighbour)

**Table 3:** Computed validation metrics for PK II dataset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Training/Test | AUC | Precision | Recall | F1 score | P-Value | Sensitivity | Specificity |
| Logistic Regression | 479/126 | 0.75 | 0.78 | 0.91 | 0.84 | 0.02 | 0.91 | 0.58 |
| Naive Bayes | 479/126 | 0.71 | 0.74 | 0.94 | 0.83 | 0 | 0.94 | 0.48 |
| K Nearest Neighbour | 479/126 | 0.70 | 0.74 | 0.95 | 0.83 | 0 | 0.95 | 0.46 |

**Table 4:** Variables used for model building for PK II

|  |  |
| --- | --- |
| **Term** | **Description** |
| ALogp2 | Square of ALogP |
| MIC3 | Modified information content index (neighborhood symmetry of 3-order) |
| PubchemFP566 | Pubchem fingerprint 566 |
| AMR | Molar refractivity |
| CrippenLogP | Crippen's LogP |
| ATSC6p | Centered Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities |
| PubchemFP431 | Pubchem fingerprint 431 |
| TopoPSA | Topological polar surface area |
| nX | Number of halogen atoms |
| minssssC | Minimum atom-type E-State: >C< |
| AATSC2v | Average centered Broto-Moreau autocorrelation - lag 2 / weighted by van der Waals volumes |
| ATSC2s | Average centered Broto-Moreau autocorrelation - lag 2 / weighted by I-state |



**Figure 6:** Comparing performance of classification algorithms on PK II

**Dataset Tox I: Classification models for prediction of drug-induced liver injury**

Size: 1104 compounds with a binary classification of human hepatotoxicity

Reference: Dataset adapted from Thakkar *et al*., 2020 (doi: [10.1016/j.jmgm.2019.107516](https://doi.org/10.1016/j.jmgm.2019.107516))

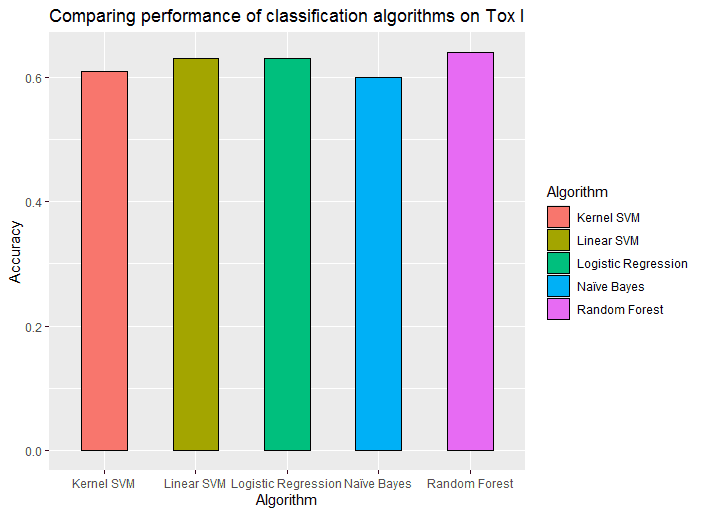
Modeling approach: Binary classification using machine learning algorithms (Logistic regression, random forest classifier, naïve bayes classifier, linear svm, kernel svm)

**Table 5:** Computed validation metrics for Tox I dataset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Training/Test | AUC | Precision | Recall | F1 score | P-Value | Sensitivity | Specificity |
| Logistic Regression | 774/330 | 0.56 | 0.67 | 0.82 | 0.74 | 0 | 0.82 | 0.29 |
| Random Forest | 774/330 | 0.58 | 0.69 | 0.79 | 0.74 | 0 | 0.79 | 0.36 |
| Linear SVM | 774/330 | 0.54 | 0.66 | 0.86 | 0.75 | 0 | 0.86 | 0.23 |
| Kernel SVM | 774/330 | 0.51 | 0.64 | 0.86 | 0.74 | 0 | 0.86 | 0.15 |
| Naive Bayes | 774/330 | 0.50 | 0.64 | 0.84 | 0.73 | 0 | 0.84 | 0.16 |

**Table 6:** Variables used for model building for Tox I dataset

|  |  |
| --- | --- |
| **Term** | **Description** |
| TopoPSA | Topological polar surface area |
| nAcid | Number of acidic groups |
| IC1 | Information content index (neighborhood symmetry of 1-order) |
| ATSC4c | Centered Broto-Moreau autocorrelation - lag 4 / weighted by charges |
| FMF | Complexity of a molecule |
| C1SP2 | Doubly bound carbon bound to one other carbon |
| ATSC5i | Centered Broto-Moreau autocorrelation - lag 5 / weighted by first ionization potential |
| ATSC2c | Average centered Broto-Moreau autocorrelation - lag 2 / weighted by charges |
| AATS0p | Average Broto-Moreau autocorrelation - lag 0 / weighted by polarizabilities |
| nRotBt | Number of rotatable bonds, including terminal bonds |



**Figure 7:** Comparing performance of classification algorithms on Tox I