Neural Net Ensemble Based QSAR Modeler for Drug Blood Brain Barrier Permeation

Mati Golani, Idit I. Golani

Abstract— **The blood-brain barrier (BBB) presents a real challenge to the pharmaceutical industry. The BBB is a very effective screener of diverse kinds of bacterial infections. Unfortunately, this functionality prevents from many drugs to penetrate it. In order to improve drug development process an assessment model is required. Effective assessment model can drastically reduce the development time, by cutting off drugs with low success rates. It also saves considerable amount of money since clinical trials focus mainly on drugs with higher likelihood of permeation.**

This work addresses the challenge by means of artificial neural net (ANN) based assessment tool. Neural network based approach is well known in the pharmacokinetic domain. In comparison with multi-linear regression, ANNs are more flexible, robust, and better at prediction. Another addressed issue is that drug data often contains correlated or skewed information. This can then lead to the construction of poor regression models.

The presented assessment tool is combined of a neural net ensemble, a group of trained neural nets that correspond to an input value set with a prediction of the barrier permeation. The returned output is the median of the ensemble's members output. The input set is composed of drug physicochemical properties such: Lipophilicity, Molecular Size (depends on Molecular Mass/Weight), Plasma Protein Binding, PSA – Polar Surface Area of a molecule, and V_d – Volume of Distribution, and Plasma Half Life (t ½).

Given the relatively small learning data-set, leave one out (LOO) which is a special case of k-fold cross validation is conducted. Although the training effort for building ANNs is much higher, in small data-sets ANNs yield much better model fitting and prediction results than the logistic regression.

Index Terms— **BBB, Pharmacokinetics, Neural net, Brain to plasma ratio.**

I. Introduction

The blood-brain barrier (BBB) presents a real challenge to the pharmaceutical industry. The BBB is very effective screener of diverse kinds of bacterial infections. Unfortunately, this functionality also prevents many drugs from penetrating it. In order to improve drug development process an assessment model is required. Effective assessment model can drastically reduce development times, by cutting off drugs with low success rates. It also saves considerable amounts of money since clinical trials focus mainly on drugs which are more likely to succeed on their task.

A. Pharmacology perspective

The Blood Brain Barrier (BBB) consists of a monolayer of brain micro vascular endothelial cells (BMVEC), which are joined together by tight junctions and form a cellular membrane [1][2]. BMVECs surrounded by a basement membrane, together with other components: pericytes, astrocytes and microglia, compose a neurovascular unit [2].

The BBB has a carrier function which is responsible for the transport of nutrients into the brain and removal of metabolites from it. While small lipid-soluble molecules (e.g. ethanol) diffuse passively through the BBB, other essential polar nutrients (glucose, amino acids) require some specific transporters. The BBB has also a barrier function that restricts the transport of potentially toxic substances through the BBB. This is achieved by a para-cellular barrier (tight endothelial junctions); trans-cellular barrier (endocytosis and trans-cytosis); enzymatic barrier (proteins with enzymatic activities) and efflux transporters. The specific barrier function of the BBB is important for preventing Central Nervous System (CNS) from harmful xenobiotics, but at the same time, prevents or limits the penetration of many drugs to the CNS [3].

The ability of these drugs to penetrate the BBB or be transported across the BBB is mainly dependent on their physiochemical properties and their affinity to a specific transport system [4].

B. Common Descriptors

Drug distribution into the CNS depends on the physicochemical properties of the compound, including: lipophilicity (logP), molecular weight (MW), and PK parameters such as: protein binding, volume of distribution (V_d), half-life etc. [5].

• Lipophilicity - Compound lipophilicity plays an important role in the absorption, distribution, metabolism, and excretion (ADME) of therapeutic drugs. Lipophilicity is often expressed as Log P, logarithm of partition coefficient P between lipophilic organic phase (1-octanol) and polar aqueous phase. While high degree of lipid solubility favors crossing the BBB by transmembrane diffusion, it also favors uptake by the peripheral tissues, thus it can lower the amount of the drug presented to the BBB [6]. In many situations lipophilicity is a good predictor of BBB penetration [7].

• Molecular weight - The optimal molecular mass for

Manuscript received July 22, 2014; revised Agust 06, 2014.

Mati Golani is with Ort Braude College, Department of Software Engineering, P.O.Box 78 Snunit 51, Karmiel 21982 Israel (phone: + 972-4-9086464; fax: +972-4- 9901-852; e-mail: matig@braude.ac.il).

Idit I. Golani is with Ort Braude College Israel, Department of Biotechnology Engineering (e-mail: igolani@braude.ac.il).

passage into the brain lies in the region of 300 to 400 Da [8][9][10] [11]. The best approximation of molecular size influence on BBB penetration is that it is inversely related to the square route of a molecular weight [12].

A limited number of drugs with high lipophilicity and low molecular size can penetrate to the brain mainly by passive diffusion.

• Polar Surface Area (PSA) - PSA is defined as the sum of polar atoms surface (oxygen, nitrogen and attached hydrogen) in a molecule. This parameter has been shown to correlate very well with BBB penetration [13][14]. BBB permeation decreases 100-fold as the surface area of the drug is increased by 2-fold (from 52 angstroms to 105 angstroms) [8].

• Protein binding - The extent of drug distribution into tissues, including the CNS, depends on the degree of plasma protein binding (albumin, α1-acid glycoprotein, and lipoproteins). Only unbound drug is available for passive diffusion through the BBB and for pharmacologic effect. The penetration rate into the brain is slow for highly protein-bound drugs [15].

• Volume of distribution (V_d) - is a proportionality factor that relates to the amount of a drug to its measured concentration. The apparent volume of distribution is a theoretical volume of fluid into which the total drug administered would have to be diluted to produce the concentration in plasma. Some drugs distribute mostly into fat, others remain in extracellular fluid, while the rest are bound extensively to specific tissues. For a drug that is highly tissue-bound, very little drug remains in the circulation, thus plasma concentration is low and volume of distribution is high [16].

• Brain/Plasma ratio (Permeation measure) - The most common method to study brain penetration in vivo is the determination of the brain/plasma ratio in rodents. For that, the test compound is dosed and both plasma and brain are sampled. The logBB describes the ratio between brain and blood (or plasma) concentrations and provides a measure of the extent of drug permeation through the BBB

$$\mathrm{LogBB} = \frac{\mathrm{AUC}_{\mathrm{tot.brain}}}{\mathrm{AUC}_{\mathrm{tot.blood}}}$$

Another in vivo measurement of CNS permeation is the log of the permeability-surface area coefficient (log PS) which is considered to be the most appropriate in vivo measurement [17][18]. However, this is a resource-intensive measure that requires microsurgical expertise. This method's advantage is by eliminating drug's serum binding. Nevertheless, by using log BB together with plasma protein binding, one can produce same or even better results.

During drug development, in vitro, ex vivo and in vivo models have been developed in order to examine the mechanisms by which different drugs penetrate the BBB.

Tissue distribution studies are commonly conducted by a traditional method using radiolabeled compounds. Brain tissue is homogenized and precipitated, and the total brain concentration of the radioactive compound is determined using liquid scintillation counting and related to its concentration in plasma.

An alternative method is quantitative microdialysis, a widely used technique that permits quantifications of drug transport to the brain. Drug concentrations measured by microdialysis are influenced by properties of the probe and perfusion solution, by the post-surgery interval in relation to surgical trauma, and tissue integrity properties [19].

All the mentioned methods for drugs permeation to the BBB are labor intensive, demand expensive compounds and equipment and use many animals. Rapid screening methods can speed up discovery and minimize the number of drug candidates for further detailed studies.

Such computational models which exist since the 1980's are based on drug's lipophilicity, hydrogen-bond potential, pKa/charge and molecular size [20][21][22][23][24]. However, in these models, other factors that can determine drug's concentration at the brain capillary surface, are not included. Factors such plasma protein binding or volume of distribution (V_d), which are present in the presented model

The rest of the paper is organized as follows: Section II introduces known structured and non-structured based modeling methods in pharmacokinetics, and present the neural-net based model. In Sections III,IV,V we present data pre-processing, training, and testing results, respectively. Finally, in Section VI, we introduce the architecture for implementation, and show results.

II. MODELING TECHNIQUES

Diverse modeling techniques such multi-linear regression, clustering, Neural nets, Bayesian neural nets [25], and decision-trees [26] where introduced with regard to pharmacokinetics modeling.

A fashionable classification of BBB permeation as appear in some published papers is to classify the BBB permeation measure into two classes: "good" (CNSp+), and "bad" (CNSp-) [27]. While the measure is indeed qualitative, a finer resolution classification may provide better comparable order between candidate drugs performance.

A neural network (ANN) is a mathematical model which is based on the biological brain structure. Interconnected processing units that form a network.

A. ANN for Pharmakokinetic Modeling

Neural network based approach is well known in the pharmacokinetic domain [27]. In comparison with multi-linear regression, ANNs are more flexible, robust, and better at prediction [28]. Furthermore, multi-linear regression is more sensitive to the relationship between the number of patterns and number of variables, thus it needs to be monitored in order to avoid chance effects [29]. Another disadvantage is that drug data often contains correlated or skewed information. This can then lead to the construction of poor regression models [28].

A distinguishing feature of neural networks is that knowledge is distributed throughout the network itself rather than being explicitly written into the program. The network then learns through exposure to diverse input set with known

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

output.

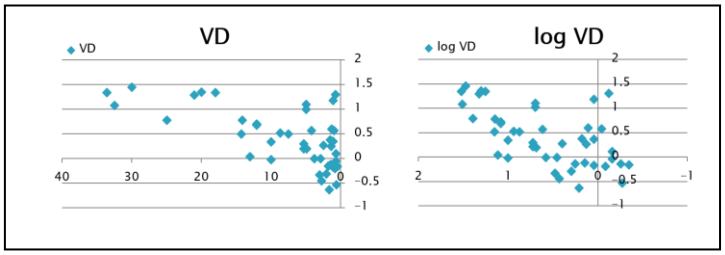

Fig. 1. Better uniformity in distribution with Log V_D

III. DATA PRE-PROCESSING

Relevant data of 47 drugs was collected from the literature. Only drugs for which all required metrics were available, were collected.

A. Consideration

Neural network training can be executed in a more efficient manner if certain preprocessing steps on the network inputs and targets are performed. Prior the network design process, the data is collected and prepared. It is generally difficult to incorporate prior knowledge into a neural network; therefore the network can only be as accurate as the data that is used for training it. After the data has been collected, there are two crucial steps to be performed before training starts: the data is uniformly distributed and then normalized.

B. Data Distribution

Some properties of the collected drugs have poor distribution. ANN prediction results tend to be more promising when the data is properly distributed. In order to improve the data distribution log operator was applied on V_d, Half Life and Brain to Plasma Ratio properties. V_d values are presented in Fig. 1. It can be concluded from Fig. 1 that input data has better distribution using a log operation with regard to V_d.

C. Normalization

Normalized data has a common base, which means that every member is evaluated for each metric with respect to other members metric in the group on a scale range of [-1,1]. ANN's perform much better on normalized data sets. The normalization step was applied on the input and the target vectors of the data set (BBB permeability).

D. Measures and precision concern

As mentioned in Section II the permeability measure should be qualitative rather than quantitative. This is due to the fuzzy nature of measurement and lack of persistence. For example, plasma to brain ratio of Clozapine (Clozaril) appears as 24 [30], or 4.1 [31], which is a large gap. Therefore it was decided that the drugs will be divided into four permeability groups in the range 1-4 where 1 is the least permeable drug, and 4 is the most. A deviation value of 1 is acceptable, and considered a success. That means that a drug that was detected as group 3 and is actually a member of group 4, will be considered a positive detection.

IV. TRAINING

The chosen ANN topology was a feed-forward back propagation network. In such topology, each input measure is connected to an input neuron; there may be one or more hidden layer neurons, and an output neuron that provides the output measure (permeability).

A. Hidden layer size

One should take into consideration when comparing networks with relatively similar accuracy, that the smaller the network, the more general it is in terms of model. When the network size increases, it may just encapsulate the specific data set instead of the general model. In order to determine the proper hidden layer size, an initial training

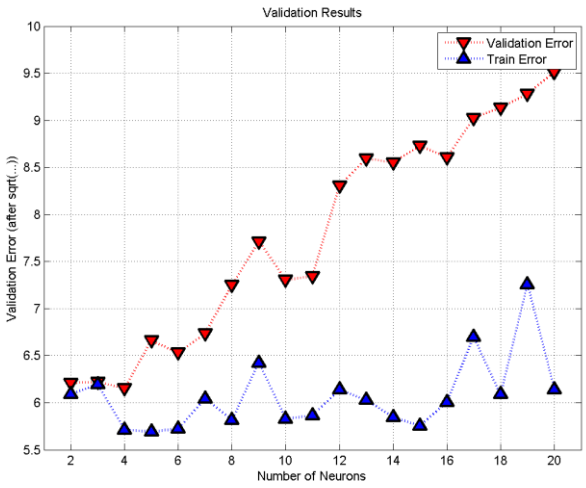

Fig. 2. Validation/training error vs. hidden layer size

phase was conducted on networks with variable hidden layer size. As reflected in Fig. 2, it infers that a hidden layer of 2 to 3 neurons provides best results. Bigger layers maybe provide better results with respect to the training error, but this result is actually misleading since it is a symptom of over-fitting. As shown in Fig. 2, beyond 3 neurons in the hidden layer, the validation error actually increases and those networks reduce generalization.

B. Early stopping

In machine learning, early stopping is a known method for improving generalization. The data is divided into training-set and validation-set.

The training set is used for computing the gradient and updating the network weights and biases. The validation set is used for monitoring. The error on the validation set is monitored during the training process. The validation error normally decreases during the initial phase of training, as does the training set error. However, when the network begins to over-fit the data, the error on the validation set typically begins to rise. When the validation error increases for a specified number of iterations, or beyond a predefined threshold –Alpha, the training is stopped. Early stopping is

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

effectively limiting the used weights in the network and thus imposes regularization.

C. Cross validation

In small data sets leave-one-out (LOO) cross-validation is normally applied. This is a special case of k-fold cross validation [27][32]. With a very small sample size (18 bankrupt and 18 non-bankrupt firms), Fletcher and Goss employ an 18-fold cross-validation method for model selection. Although the training effort for building ANNs is much higher, ANNs yield much better model fitting and prediction results than the logistic regression [33].

D. Net tournament

During the cross validation, and for each fold, a tournament between 100 networks was conducted. Only the winner network with best results during this fold (minimal error) was retained, as illustrated in Fig. 3.

E. ANN ensemble

Since a network training tournament is performed for each fold, the outcome is a group of winning ANN's, one for each sample. Most often one would pick the best performing network. Nevertheless, here we suggest a different approach, i.e. a neural net ensemble. A neural net ensemble is a group of ANN's that provides a single output to a given input. This output can be the average of the ensemble members output, a quorum based result, median, etc. In this work we have used the median of ensemble members output as the ensemble's output.

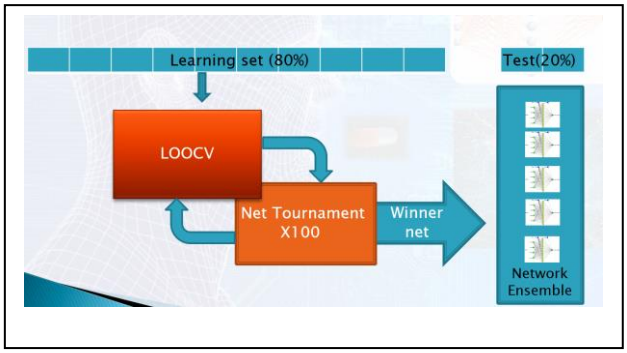

Fig. 3. design structure of the learning phase and ensemble generation

V. TESTING

- The preprocessed data was repetitively (several hundred test cycles), and randomly divided into a Training (80%) and Testing (20%) groups
- The training set was utilized to generate the ANN ensemble.
- Drug data from the test group was presented to the ensemble, and its output was compared to the known one- in terms of permeation group membership.
- Results were grouped with accordance to the delta between the predicted and actual permeability group (see Section III.D)

The test phase is illustrated in Fig. 4. Fig. 6 provides a graphical presentation of the results. The presented results are the average of several hundred runs. In addition, diverse early stopping settings were investigated it terms of maximal number of epochs, and alpha. Most combinations provided similar and satisfactory results of 89% success rate.

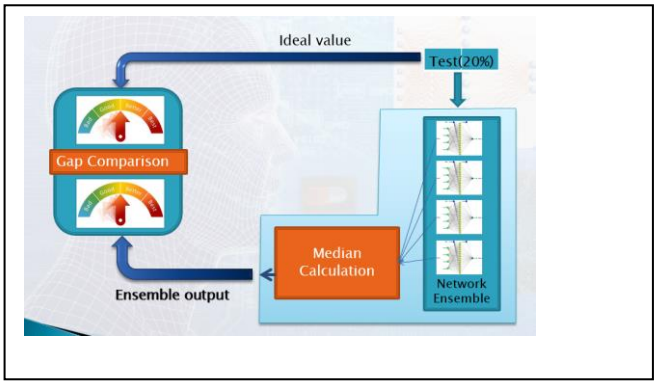

Fig. 4. Test phase

VI. ARCHITECTURE

The first analysis to determine the hidden layer size was conducted both on Matlab and the Encog framework.

Next phases as described in Section V, were implemented in C# using the Encog .NET package. Data was saved on a server database. The application was designed as a client/server application. An illustrative screenshot of the training phase is presented in Fig. 5.

VII. CONCLUSION

In this paper, we propose a new ensemble neural net based mechanism as an evaluator for drug-BBB permeation. Design time evaluator for drug development – in particular, by providing finer permeation classification, with relatively high success rates. Due to the small given data set, a leave one out cross validation technique was performed.

Our goal is to develop an approach that allows an interactive drug design that is less labor intensive, or demand expensive with respect to compounds/equipment, thus uses less animals. This can be achieved by using such a mechanism for scoring candidates, and performing the more expensive pre-clinical stages on the provided best candidates.

Our specific contribution in this work is twofold. We have incorporated plasma protein binding as a parameter into the model, and we also propose the modeling mechanism that provides finer permeation resolution while coping with relatively smaller data sets. The benefits of this approach have been discussed in this work.

One main avenue for a future research involves larger dataset incorporation. We would also like to extend our model with metrics such as: Hydrogen-bonding (Hydrogen bond acceptor/donor), and drug's affinity to efflux transporters such as P-glycoprotein (P-gp).

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

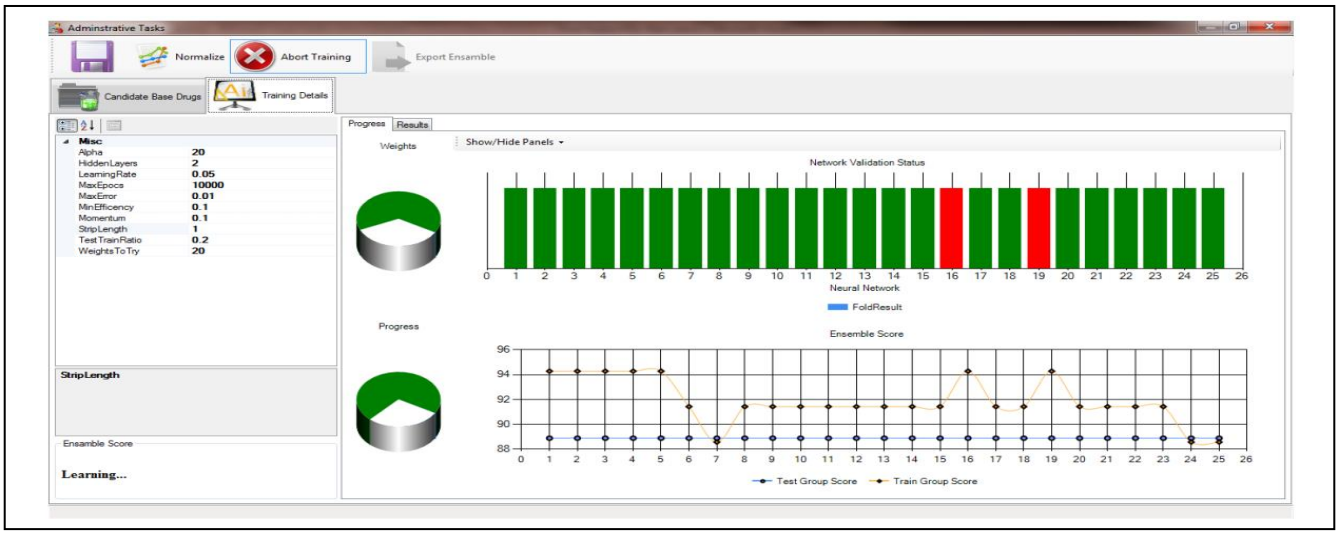

Fig. 5. Training phase UI

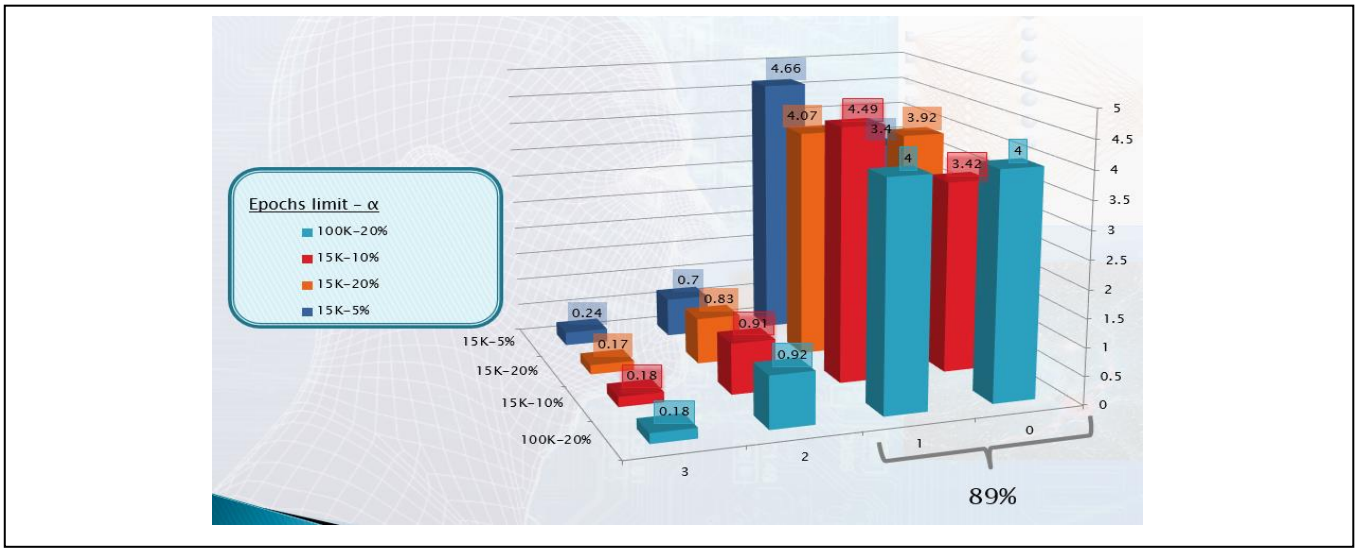

Fig. 6. Permeability group detection precision

REFERENCES

[1] N.J. Abott, "Physiology of the blood-brain barrier and its consequences for drug transport to the brain". International Congress Series. 1277, 3-18, 2005.

[2] F.L. Cardoso, D. Brites, and M.A. Brito, "Looking at the blood-brain barrier:Molecular anatomy and possible investigation approaches", 2010. Brain Research Reviews. 64:328-363

[3] W. Loscher, and H. Postschka, "Role of drug efflux transporters in the brain for drug disposition and treatment of brain diseases". Progress in Neurobiology, 2005. 76:22-76

[4] DJ. Begley, "ABC transporters and the blood-brain barrier". Curr Pharm Des 2004. 10(12):1295-312

[5] S. Schmidt, D.Gonzalez, and H. Derendorf, "Significance of protein binding in pharmacokinetics and pharmacodynamics". J Pharm Sci. 99(3):1107-22, 2010.

[6] NH Greig, A. Brossi, XF Pei, DK Ingram, and TT. Soncrant, " Designing drugs for optimal nervous system activity". In New Concepts of a Blood-brain Barrier. Edited by Greenwood J, Begley DJ, Segal MB. New York: Plenum Press; 1995::251-264

[7] RN Waterhouse, "Determination of lipophilicity and its use as a predictor of blood-brain barrier penetration of molecular imaging agents". Mol Imaging Biol. 2003, 5(6):376-89.

[8] H. Fischer, R. Gottschlich, and A. Seelig, "Blood-brain barrier permeation molecular parameters governing passive diffusion". J Membr Biol , 1998, 165: 201–211.

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[9] WM. Pardridge, “Brain drug targeting: the future of brain drug development”. Cambridge, UK: Cambridge University Press, 2001.

[10] VA Levin, “Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability”. J Med Chem 1980, 23:682-684.

[11] H. van de Waterbeemd, G. Camenisch, G. Folkers, JR. Chretien, and OA. Raevsky, “Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors”. J Drug Target 1998 6: 151–165

[12] K. Felgenhauer, “Protein filtration and secretion at human body fluid barriers”. Pflugers Arch. 1980 384:9-17.

[13] S. Shityakov, W. Neuhaus, T. Dandekar, C. Förster, “Analysing molecular polar surface descriptors to predict blood-brain barrier permeation”. Int J Comput Biol Drug Des. 2013 6(1-2):146-56

[14] J. Kelder, PD. Grootenhuis, DM. Bayada, LP Delbressine, JP. Ploemen, “Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs”. Pharm Res. 1999 16(10):1514-9

[15] R. Nau, F. Sorgel, and H. W. Prange, “Lipophilicity at pH 7.4 and molecular size govern the entry of the free serum fraction of drugs into the cerebrospinal fluid in humans with uninflamed meninges”. J. Neurol. Sci. 1994, 122:61-65

[16] H. van de Waterbeemd, “Which in vitro screens guide the prediction of oral absorption and volume of distribution?”, Basic Clin Pharmacol Toxicol 96 (2005): 162–166

[17] I. Martin, “Prediction of blood-brain barrier penetration: are we missing the point? “Drug Discov Today 2004, 9:161-162.

[18] WM. Pardridge, “Log(BB), PS products and in silico models of drug brain penetration”. Drug Discov Today 2004, 1;9(9):392-3.

[19] E.C. De Lange, and M. Danhof, “Considerations in the use of cerebrospinal fluid pharmacokinetics to predict brain target concentrations in the clinical setting: implications of the barriers between blood and brain”. Clin. Pharmacokinet. 2002, 41:691-703.

[20] RC. Young et al., “Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists”. J Med Chem. 1988, 31(3):656-71.

[21] SG. Jezequel, “Central nervous system penetration of drugs: importance of physicochemical properties”. Progr Drug Metab; 1992, 13: 141–178

[22] EG. Chikhale, KY. Ng, PS Burton, and RT. Borchardt, “Hydrogen bonding potential as a determinant of the in vitro and in situ blood-brain”, Pharm Res 1994, 11(3):412-9

[23] F. Atkinson, S. Cole, C. Green, and H. van de Waterbeemd, “Lipophilicity and Other Parameters Affecting Brain Penetration”, Curr Med Chem – CNS Agents 2002, 2(3) 229-240

[24] MH. Abraham, “The factors that influence permeation across the blood-brain barrier”, Eur J Med Chem. 2004, 39(3):235-40.

[25] JT. Goodwin , DE. Clark, “In silico predictions of blood-brain barrier penetration: considerations to "keep in mind"”. J Pharmacol Exp Ther., 2005,;315(2):477-83.

[26] C. Suenderhauf, F. Hammann, and J. Huwyler, ”Computational Prediction of Blood-Brain Barrier Permeability Using Decision Tree Induction”. Molecules 17(9) 2012, 10429-10445

[27] J.V. Turner, D.J. Maddalena, and D.J Cutler, “Pharmacokinetic parameter prediction from drug structure using artificial neural networks”, International Journal of Pharmaceutics, Volume 270, Issues 1–2, 11 February 2004, Pages 209-219

[28] D. Butina, MD. Segall, and K. Frankcombe, “Predicting ADME properties in silico: Methods and models”, Drug Discovery Today, 2002, 7: S83-S88.

[29] JG. Topliss , RP. Edwards, “Chance factors in studies of quantitative structure-activity relationships”. Journal of Medicinal Chemistry, 1979, 22: 1238-1244.

[30] G. Zhang, A. Terry Jr, and M.G. Bartlett, “Sensitive liquid chromatography/tandem mass spectrometry method for the simultaneous determination of olanzapine, risperidone, 9-hydroxyrisperidone, clozapine, haloperidol and ziprasidone in rat brain tissue”. Journal of Chromatography B, (2007) 858(1), 276-281.

[31] T.S. Maurer, D.B. DeBartolo, D.A. Tess, and D.O. Scott, “Relationship between Exposure and nonsprcific binding of thirty-three Central Nervous System Drugs in mice”, Drug Metabolism and Disposition volume 33(1), 2005,pages 175-181.

[32] I.V. Tetko, D.J. Livingstone, and A.I. Luik, "Neural Network Studies. 1. Comparison of Overfitting and Overtraining," J. Chem. Info. Comp. Sci., 35, (1995), 826-833.

[33] D. Fletcher, E. Goss, “Forecasting with neural networks: An application using bankruptcy data”, Information and Management 24 (1993) 159 -167.

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)