In silico modelling of chylomicron association to predict lymphatic absorption of small molecules

Jed Malec[1], Jong Bong Lee[2], Atheer Zgair[2], Beth Williamson[1], Pavel Gershkovich[2]

[1]DMPK, Evotec, Milton Park, Abingdon, UK

[2]School of Pharmacy, University of Nottingham, Nottingham, UK

evotec

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Background & Objectives

Intestinal lymphatic transport

The lymphatic system displays a range of important functions in the body and has key roles in many diseases, such as infections, cancer and metastasis, immune and inflammatory conditions and metabolic diseases, amongst others[1]. Recent advances in these areas have increased the interest in targeted delivery to the lymphatic system. Association of drugs with chylomicrons (CM) in the enterocytes is a key process for intestinal lymphatic transport. However, early prediction of intestinal lymphatic transport is difficult due to limited *in vitro* tools. *In vivo* measurements are invasive, difficult to perform and expensive. *Ex vivo* measurements of association of drugs with CM have proved to be a reliable method for predicting intestinal lymphatic transport[2,3], though its throughput is limited.

Objectives

To provide a quick and 'easy-to-use' tool for the prediction of intestinal lymphatic transport.

Lymphatic transport

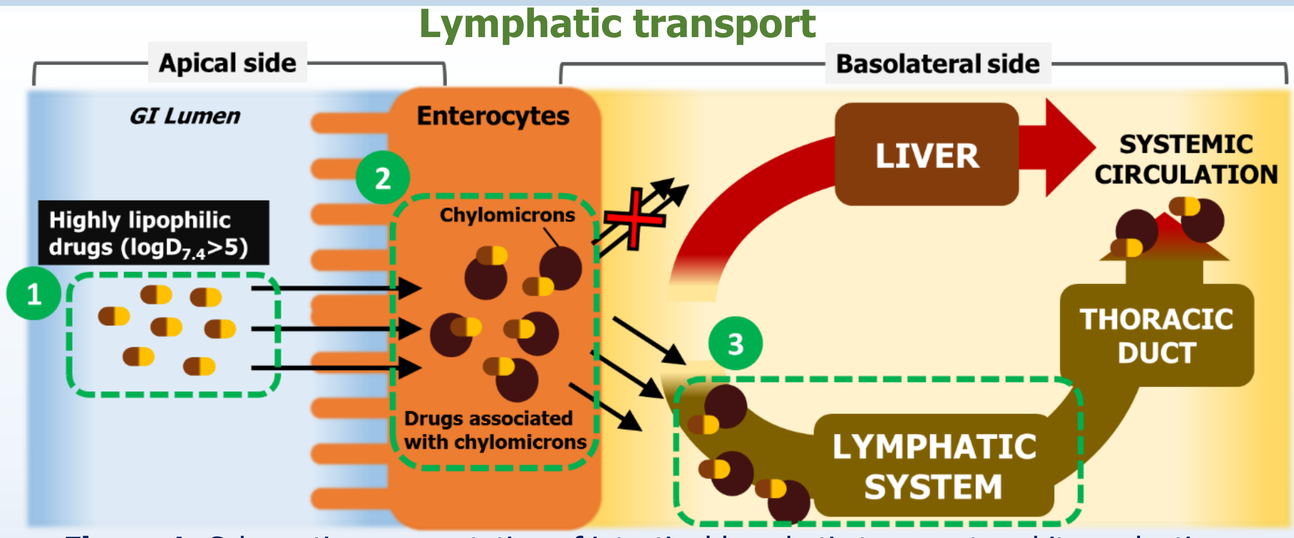


Figure 1: Schematic representation of intestinal lymphatic transport and its evaluation.

We have utilised literature data and that generated in our labs to create a predictive *in silico* model, based on Random Forest algorithm. We have evaluated different models and analysed the influence of descriptors to gain further insight into physico-chemical properties that govern CM association, a process not yet described in detail.

Methods

Data set & descriptors

- Data obtained from peer-reviewed articles[2,3] and our laboratory employing methods established in literature[2].
- Data set of 61 compounds: 20 cannabinoids (CBDs), 25 bexarotene (BEX) prodrugs and 16 literature molecules.
- 2 sets used for modelling: 1) whole data set, and 2) CBDs and BEX prodrugs only.
- 13 descriptors in total - 10 basic physico-chemical properties (ACD/Labs) and 3 descriptors obtained by combination of the others: ionisation in water (LogP-LogD), $LogD_{7.4}$ divided by number of heavy atoms (LogD/HA) and polar surface area normalised to molecular volume (PSA/MV).
  - The last two represent different measurements of lipophilicity and allow for better differentiation of structurally similar molecules.
- There is no correlation between descriptors ($r^2 < 0.75$ in all cases).

Model development

- Modelling was completed in R software[4], using packages Caret[5] and Random Forest[6] (RF).
- Initial model was developed using all descriptors. After assessing descriptor importance, using RF built-in feature in R, they were systematically filtered by removing those with lowest importance, while maintaining model's accuracy.
- Descriptor importance was also assessed using recursive feature elimination, and the 6 most important descriptors were in the same order as when using the RF built-in feature.
- Data was normalised using $LogK^x$: $\boldsymbol{LogK = Log(^{1-x}/_{x})}$
- The error of the model was measured by means of root mean square error (RMSE) of the Out-of-Bag (OOB) predictions in LogK scale, which correlates with cross-validation. Geometric-mean fold deviation (GMFD) was also calculated.
- Optimisation of the *mtry* parameter showed the default value gave the best prediction.
- All models were completed using 500 trees, where *ntrees* > ~200 were equivalent.
- The final model was established using 6 descriptors, as shown in Figure 4.

Results

Fig. 1

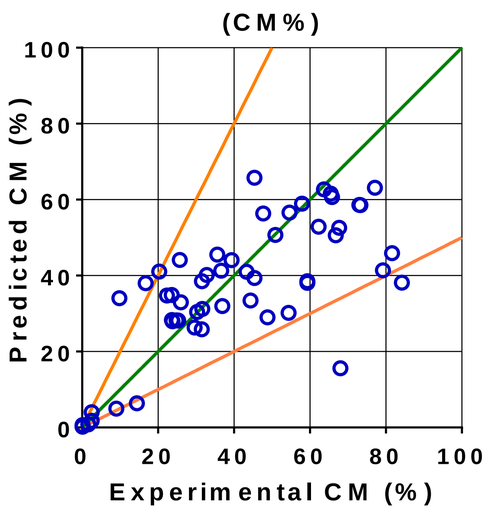


Fig. 2

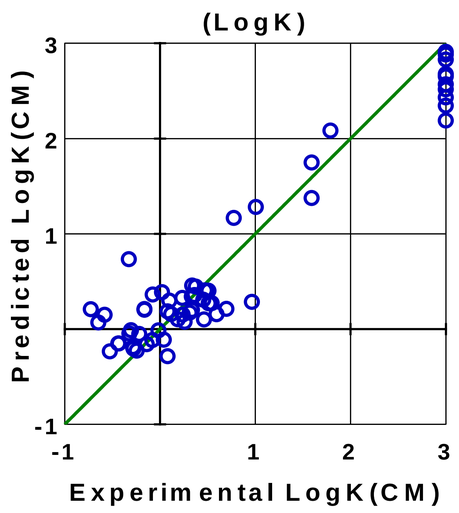


Fig. 3

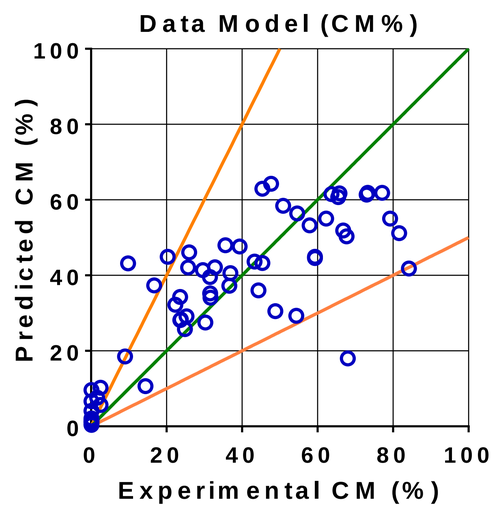


Fig. 4

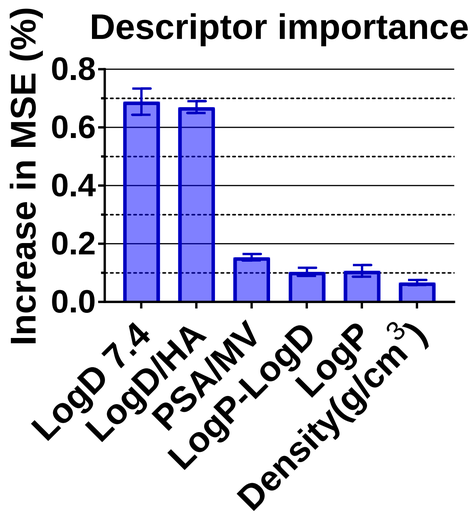
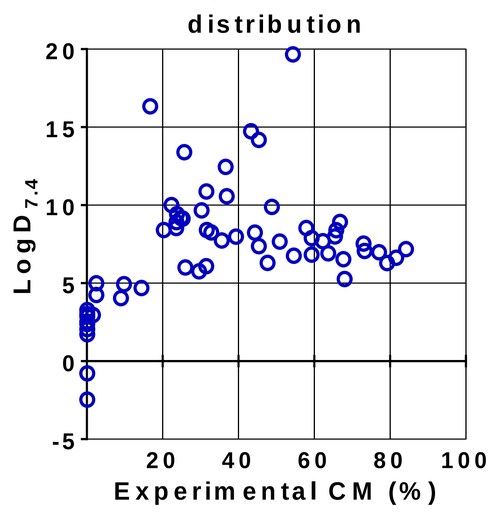


Fig. 5



Figures 1 and 2 show the results for the model developed with the whole data set (61 compounds). The data is presented as % CM association (Fig. 1) and as normalised by LogK (Fig. 2). For all graphs, the data points are represented by blue circles, the green line is unity, while the orange lines represent 2-fold error. Few compounds are predicted with deviation higher than 2-fold, the biggest outlier being halofantrine. Its physico-chemical properties (relatively low lipophilicty) do not suggest high CM association, as predicted by the model. The high measured value requires further investigation. Figure 3 shows the results for a model developed with non-normalised data. LogK normalisation helped decrease the over-prediction of compounds with low CM association. The RMSE of the non-normalised model for CM% > 10% was the same as for the normalised model.

The descriptors used in the final model and their importance can be seen in Figure 4. The error bars represent standard deviation of the importance value for the model repeated using three different *seed* values. Lipophilicity is the biggest driver for the CM association according to the model. However, as it can be seen in Figure 5, LogD alone correlates poorly with CM association and is not sufficient to explain the different levels of association.

Table 1

	Whole data set model	CBDs & BEX prodrugs only (46 compounds)		Whole data set model for >10% CM and without Halofantrine
		Whole data model	CBD & BEX prodrugs only model	
RMSE	0.36	0.29	0.34	0.30
GMFD	1.49	1.35	1.44	1.30

Table 1 shows the accuracy of both models expressed as RMSE and GMFD. Establishing a model with only CBDs and BEX prodrugs resulted in a larger error when compared to the predictions from the whole data model. Molecules with low CM association were predicted with a larger RMSE, but, since we're interested in design of molecules with high association, we assessed the RMSE for molecules with >10% CM association. The model showed good accuracy for such compounds, with low RMSE after excluding halofantrine.

Conclusions

- The model showed good predictive performance, proving to be a useful *in silico* tool for the design of compounds.
- According to the model, lipophilicity is the main factor driving CM association. The descriptors proposed in this work have proved useful, though they may not describe all the properties that govern association with CM, as some outliers are present.
- Halofantrine is the biggest outlier in this model; its relatively low lipophilicity would not suggest high CM association. This outlier requires further investigation.
- The model is limited by various factors: the chemical space covered in the data set, which may affect the importance of descriptors and its ability to predict chemically different scaffolds; the ability of RF algorithm to train this data set; and the employed descriptors, which may not be able to fully describe CM association.
- Future work will aim to evaluate the applicability of the model and the limitations mentioned above. The data set, together with efforts to analyse and improve this work, will be presented for publication in a peer-reviewed journal.

Bibliography

1. Trevaskis, N., Kaminskas, L. & Porter, C. *Nature Reviews Drug Discovery* 14, 781-803 (2015).
2. Gershkovich, P. & Hoffman, A. *European Journal of Pharmaceutical Sciences* 26, 394-404 (2005).
3. Gershkovich, P., Qadri, B., Yacovan, A., Amselem, S. & Hoffman, A. *European Journal of Pharmaceutical Sciences* 31, 298-305 (2007).
4. R development Core Team (2008). R Foundation for Statistical Computing,. www.R-project.org
5. Kuhn, M. Caret package. *Journal of Statistical Software* 28 (2008).
6. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.