Predicting Skin Permeation Using the Huskin Database

Introduction and Motivation

The skin serves as the body's outermost barrier, making it the first point of contact for many of the external molecules humans are exposed to. To regulate the movement of molecules into and out of the body, the skin has developed remarkable barrier properties. However, despite this barrier, some molecules are still able to permeate the skin. While this can pose risks, it also presents opportunities, particularly in the context of delivering pharmaceutical agents.

Considerable research has been dedicated to understanding the permeation properties of various molecules through the skin. Experimental methods have been central to these efforts, but they come with significant costs in terms of time, money, and the potential dangers associated with handling skin-permeating molecules.

In light of these challenges, computational methods for predicting skin permeability offer an attractive alternative. These methods can be faster, more cost-effective, and safer. Computational approaches to skin permeation prediction can be broadly categorized into two types: (i) physics-based methods that utilize simulation techniques (for example: [1–6, 10]) and (ii) statistical modeling that relies on existing data (for example: [7, 9, 12, 14]). Physics-based methods have the advantage of not requiring pre-existing data, and they can offer novel insights into the permeation mechanism at the (sub)-molecular level, depending on the simulation's resolution. However, these methods are computationally intensive, which can undermine the time and cost benefits of computational approaches.

On the other hand, statistical methods use comparatively lightweight parameterized models built from existing data. The recent introduction of the Huskin skin permeability database [8] provides a valuable resource for applying statistical methods to predict skin permeation. This database is particularly noteworthy for its large size and its exclusive focus on human skin permeation data.

This is not the first attempt to use this dataset for skin permeation prediction. However, this is not the first attempt to use this dataset for skin permeation prediction. Previously, Quah and colleagues [11, 13] developed statistical models based on this data. In their first study [11] , they sought to parameterize a linear model that relates a molecule's logP, topological polar surface area (TPSA), and molecular volume (MV) to the skin permeation constant ($logK_p$). Their approach involved segmenting the database into smaller subsets to build several QSAR models. This segmentation was necessary due to the variability in the Huskin database, which includes different skin sources, skin layers, temperatures, pH values, and concentrations, among other factors. The best model presented by Waters et al. demonstrated excellent R^2 and RMSE values (0.8949 and 0.35, respectively, for the test set) but was based on a relatively small dataset, with 29 data points in the training set and 7 in the test set. A larger QSAR model, which included a single entry for each molecule in the database with non-missing data (n=214), resulted in an R^2 of 0.5057 and an RMSE of 0.84 for the test set. In their second paper [13], they performed a similar analysis, but this time they used the count of functional groups / fragments to predict skin permeability.

This study aims to create a 'middle ground' dataset, larger than the subsets used by Waters et al. but more selectively filtered than simply including every molecule. Unlike Waters et al., this work will not filter based on temperature and pH; instead, these variables will be used as predictive features. Additionally, this study will explore the development of more complex models beyond the linear models employed by Waters et al.

The detailed methods and results of this work are documented in the accompanying Jupyter notebook.

References

(1) S. Yang, R. Wang, G. Wan, Z. Wu, S. Guo, X. Dai, X. Shi and Y. Qiao, *Journal of Chemical Information and Modeling*, 2016, **56**, Publisher: American Chemical Society, 2234–2242.

(2) K. Gajula, R. Gupta, D. B. Sridhar and B. Rai, *Journal of Chemical Information and Modeling*, 2017, **57**, Publisher: American Chemical Society, 2027–2034.

(3) P. Rocco, F. Cilurzo, P. Minghetti, G. Vistoli and A. Pedretti, *European Journal of Pharmaceutical Sciences*, 2017, **106**, Publisher: Elsevier, 328–335.

(4) A. Del Regno and R. Notman, *Physical Chemistry Chemical Physics*, 2018, **20**, Publisher: Royal Society of Chemistry, 2162–2174.

(5) M. Lundborg, C. L. Wennberg, A. Narangifard, E. Lindahl and L. Norlén, *Journal of Controlled Release*, 2018, **283**, Publisher: Elsevier B.V., 269–279.

(6) R. Gupta, B. S. Dwadasi, B. Rai and S. Mitragotri, *Scientific Reports 2019 9:1*, 2019, **9**, Publisher: Nature Publishing Group, 1–11.

(7) A. Burli, R. M. Law, J. Rodriguez and H. I. Maibach, *Regul Toxicol Pharmacol*, 2020, **112**, 104614.

(8) D. Stepanov, S. Canipa and G. Wolber, *Sci Data*, 2020, **7**, Number: 1 Publisher: Nature Publishing Group, 426.

(9) H. S. Cheruvu, X. Liu, J. E. Grice and M. S. Roberts, *Data Brief*, 2022, **42**, 108242.

(10) M. Lundborg, C. Wennberg, J. Lidmar, B. Hess, E. Lindahl and L. Norlén, *Biophysical Journal*, 2022, Publisher: Cell Press, DOI: 10.1016/J.BPJ.2022.09.009.

(11) L. J. Waters and X. L. Quah, *Sci Data*, 2022, **9**, 584.

(12) Y.-W. Wu, G. H. Ta, Y.-C. Lung, C.-F. Weng and M. K. Leong, *Pharmaceutics*, 2022, **14**, 961.

(13) L. J. Waters, D. J. Cooke and X. L. Quah, *Sci Data*, 2023, **10**, Publisher: Nature Publishing Group, 821.

(14) K. Ita and J. Prinze, *Journal of Drug Targeting*, 2024, **32**, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/1061186X.2023.2284096, 57–65.