Supporting Information

for

Automated selection of compounds with physicochemical properties to maximize bioavailability and druglikeness

*Taiji Oashi, Ashley L. Ringer, E. Prabhu Raman, and Alexander D. MacKerell Jr.**

Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, 20 Penn Street, Baltimore, MD 21201

Corresponding author phone: (410) 706-7442; fax: (410) 706-5017

email: alex@outerbanks.umaryland.edu

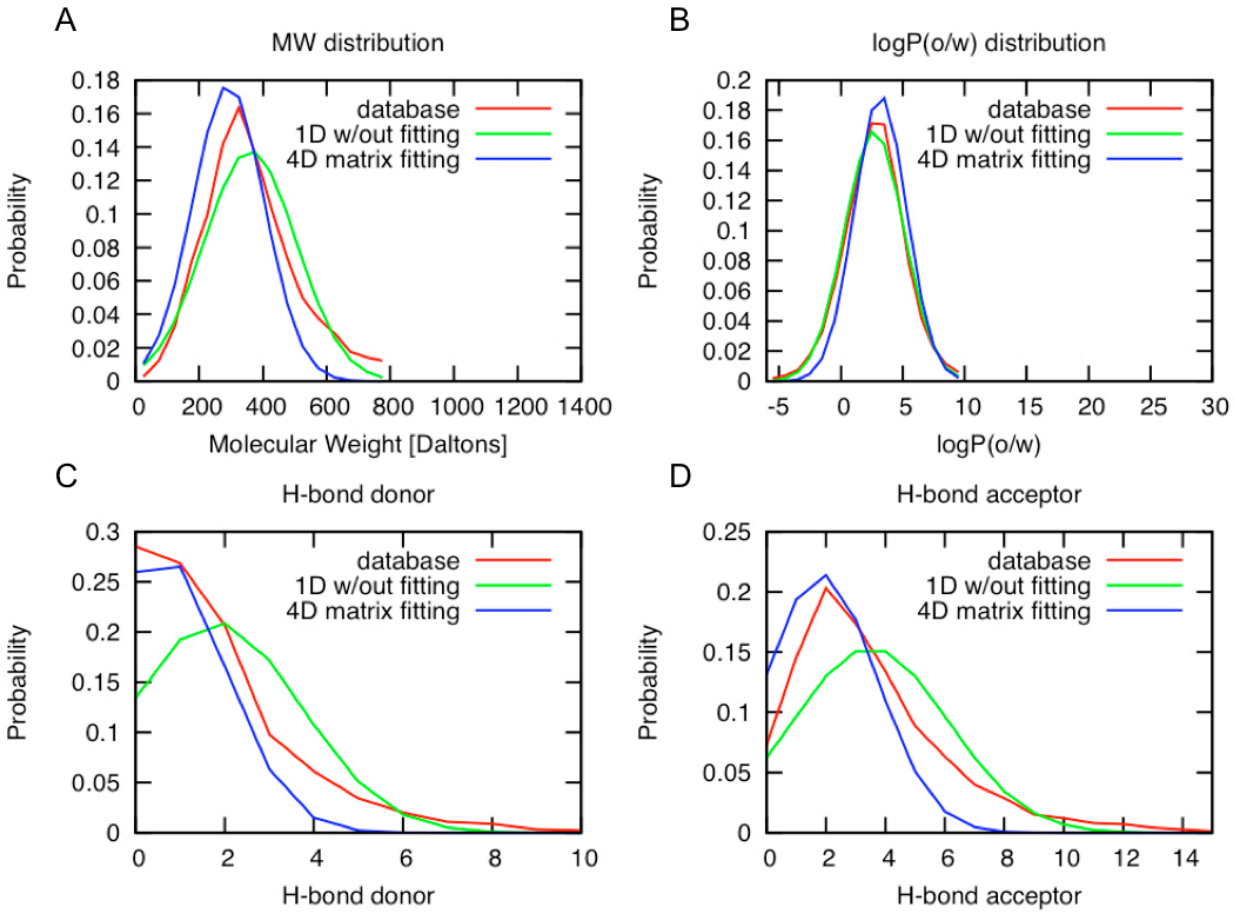

Figure S1. 1D probability distributions of physicochemical properties. (A) MW, (B) logP(o/w), (C) HDO, (D) HAC Red lines represent the distribution of the WDI database. Green lines represent the distributions based directly on the means and standard deviations from the WDI database: the distribution is not well reproduced, and especially the highest-scoring points are consistently missed. Blue lines represent the distributions of the 4D dependent multivariate model with optimized mean values and variance-covariance matrix. The results show that 4D dependent model reasonably

reproduces the distribution of the original database.

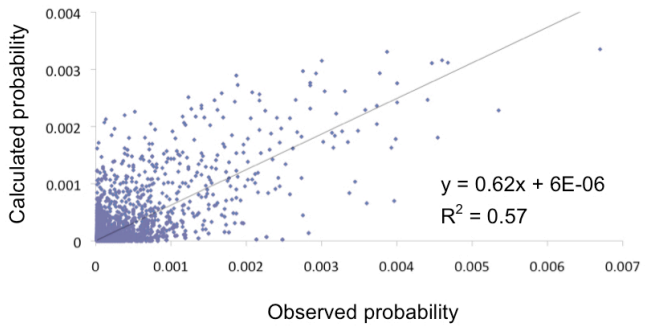

Figure S2. Comparison between observed probability from the WDI database and calculated probability based on improved 4D independent model. Linear regression analysis was performed to obtain linear function: $y = 0.62x + 6.0\times10^{-6}$. Correlation of determination R^2 is 0.57.

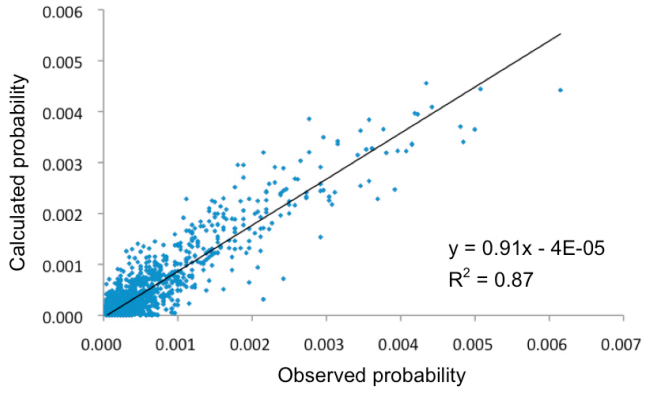

Figure S3. Comparison between test set observed probability and training set predicted probability. The entire WDI dataset were divided randomly into equal sized training and test sets. Optimized means values and variance-covariance matrix were obtained based on training set using improved 4D multivariate dependent Gaussian model, and applied to calculate probabilities using test set, that then be compared with observed test set probabilities. The analysis was performed to evaluate the possibility of

overfitting. Linear regression analysis was performed to obtain linear function: $y = 0.91x - 4.0\times10^{-5}$. Correlation of determination R^2 is 0.87. The result shows that the model with R^2 value of 0.89 in Figure 4 was not overfitted, thereby the model would be robust and used to estimate 4D-BA values for new compound sets that facilitates drug discovery process.

Supplementary data for Figure S3

$$\mu_{initial} = \begin{pmatrix} 2.7 \\ 360.6 \\ 3.5 \\ 1.7 \end{pmatrix}, \quad S_{initial} = \begin{pmatrix} 5.9 & 129.0 & -1.2 & -1.2 \\ 129.0 & 20946.9 & 237.5 & 109.4 \\ -1.2 & 237.5 & 6.8 & 3.2 \\ -1.2 & 109.4 & 3.2 & 3.2 \end{pmatrix}$$

$$\mu_{optimized} = \begin{pmatrix} 3.2 \\ 291.8 \\ 1.9 \\ 0.6 \end{pmatrix}, \quad S_{optimized} = \begin{pmatrix} 4.4 & 126.4 & -0.9 & -0.9 \\ 126.4 & 12692.8 & 105.5 & 36.0 \\ -0.9 & 105.5 & 3.5 & 1.7 \\ -0.9 & 36.0 & 1.7 & 2.0 \end{pmatrix}$$

$\mu_{initial}$ and $S_{initial}$ were calculated based on training set. $\mu_{optimized}$ and $S_{optimized}$ were obtained via fitting using training set, that were then used to calculate 4D-BA for test set to be compared with test set observed probability to evaluate the possibility for overfitting.