Knowledge-Based Methods to Train and Optimize Virtual Screening Ensembles

Robert V. Swift[†], Siti A. Jusoh[‡,§], Tavina L. Offutt[†,§], Eric S. Li[†], Rommie E. Amaro[*,†]
[†]Department of Chemistry and Biochemistrsy, University of California San Diego. La Jolla, California, 92093-0340
[‡]Faculty of Pharmacy, Universiti Teknologi MARA Malaysia, 42300 Bandar Puncak Alam, Malaysia

[§]These two authors contributed equally to the production of this work
*Corresponding author: ramaro@ucsd.edu

Supporting Information

Standard Error of the AUC

As given in equation 3 of the paper, the AUC can be expressed by averaging TPF values determined at each inactive compound in the ranked list. Equivalently, the AUC can be determined from an average of the FPF values determined at each active compound in the ranked list; i.e. $AUC = 1 - \langle FPF \rangle_A$. Since the AUC can be determined by averaging over both active and inactive compounds, the numbers of each will contribute to the errors, and each average will have its own distribution. The standard error given in equation 5 of the paper, labeled equation S1 here in the supplementary information, incorporates both error sources.

$$SE = \sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_I^2}{N_I}} \tag{S1}$$

The variance due to active and inactive compounds are given as σ_A^2 and σ_I^2, respectively, while the number of active and inactive compounds included in the estimate are given N_A and N_I, respectively.

The variance due to actives is given by equation S2

$$\sigma^2{}_A = \langle (FPF - \langle FPF \rangle_A)^2 \rangle_A \tag{S2}$$

The *A* subscripts instruct that the averages should be carried out using FPF values determined at each active compound in the ranked list. Similarly, to determine the contribution from the inactive compounds, the following equation is used.

$$\sigma^2{}_I = \langle (TPF - \langle TPF \rangle_I)^2 \rangle_I \tag{S3}$$

Standard Error of the ROC Enrichment Factor

The value of the ROC enrichment, equation 4 of the paper, is dependent on the active compounds, through the TPF, and the inactive compounds through the FPF. As a result, error arises from both active and inactive compounds. The standard error can be derived[1] and takes the form giving in equation S4.

$$SE = \frac{1}{FPF}\sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_I^2}{N_I}} \tag{S4}$$

Similarly to equation S1, the variances of the active and inactive compounds are given σ_A^2 and σ_I^2, respectively, while the number of active and inactive compounds included in

the estimate are given N_A and N_I, respectively. The FPF value is the value at which the EF is determined. The variance due to the active compounds is given by equation S5.

$$\sigma^2{}_A = \frac{1}{FPF^2}\left(\frac{TPF(1-TPF)}{N_A}\right) \quad \text{(S5)}$$

Similarly, the variance due to the inactive compounds is given by equation S6

$$\sigma^2{}_I = \frac{1}{FPF^2}S^2\left(\frac{FPF(1-FPF)}{N_I}\right) \quad \text{(S6)}$$

In equation S6, S^2 is the square of an approximation to the slope of the ROC curve, *S*, tangent to the point where the EF value was determined. The approximation is derived from an analytic estimate of the ROC curve due to Hanley[2], $Y = X^{(1-AUC)/AUC}$, as described by Nichols[1].

$$S = EF\left(1 + \frac{log(EF)}{log(FPF)}\right) \quad \text{(S8)}$$

Training Method Scaling

Exhaustive training. For *N* conformations, the exhaustive method forms all possible ensembles at each ensemble size from 1 to *N*. For an ensemble size of *k*, with 1< *k* <*N*, the number of ensembles that can be constructed is given by the binomial coefficient,

$$\binom{N}{k} = \frac{N!}{k!(N-k)!} \quad \text{(S8)}$$

The total number of ensembles constructed, *T,* can be determined by summing the values of the binomial coefficient from 1 to *N*.

$$T = \sum_{k=1}^{N}\binom{N}{k} \quad \text{(S9)}$$

Equation S9 can be simplified by writing the binomial formula as follows.

$$(x+y)^N - x^N = \sum_{k=1}^{N}\binom{N}{k}x^{N-k}\,y^k \quad \text{(S10)}$$

If we set both *x* and *y* to a value of 1 in equation S10 and compare the results to S9, it follows that the total number of constructed ensembles grows exponentially with the number of conformations, as described by equation S11. This growth can be expressed using big O notation as, $O(2^N)$.

$$T = 2^N - 1 \quad \text{(S11)}$$

Slow heuristic training. If there are *N* conformations, in the first step of the slow heuristic method, *N* one-membered ensembles are considered, and the best performer is retained. In the second step, *N – 1* two-membered ensembles are considered and the best performer is retained. This process is repeated until a *1 N*-membered ensemble is determined. The total number of ensembles constructed, *T*, is given by equation S12.

$$T = N + N - 1 + N - 2 + \cdots + 1 = \sum_{k=1}^{N} N - (k-1) \quad \text{(S12)}$$

Equation S12 can be re-written as equation S13.

$$T = N(N+1) - \sum_{k=1}^{N} k \quad \text{(S13)}$$

The sum in the second term of S13 is known as a triangular number and can be re-written as $N(N+1)/2$, and equation S13 can be simplified.

$$T = \frac{N(N+1)}{2} \quad \text{(S14)}$$

In the limit of large N, S14 approaches $N^2/2$. Using big O notation, this is expressed as $O(N^2)$.

References

(1) Nicholls, A., Confidence Limits, Error Bars and Method Comparison in Molecular Modeling. Part 1: the Calculation of Confidence Intervals. *J Comput.-Aided Mol. Des.* **2014**, *28*, 887-918.
(2) Hanley, J. A.; McNeil, B. J., The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*, 29-36.