

Supplementary Methods

Transformation of Raw scores to Z-scores. A raw score was transformed into a z-score by taking its difference to the raw score expected at random for that combination of set sizes and dividing the result by the random standard deviation:

Given,

$$rs(S_1, S_2) = \text{raw score of set } S_1 \text{ vs. set } S_2$$

$$n(S_1, S_2) = \text{size}(S_1) \times \text{size}(S_2)$$

$$\mu(x) \approx (4.24 \times 10^{-4})x \quad [\text{Expected raw score mean, Supplementary Table 5}]$$

$$\sigma(x) \approx (4.49 \times 10^{-3})x^{0.665} \quad [\text{Expected raw score std. dev., Supplementary Table 5}]$$

Then,

$$z = (rs(S_1, S_2) - \mu(n(S_1, S_2))) / \sigma(n(S_1, S_2))$$

For the comparison of the sets Dihydrofolate reductase inhibitor (DHFR) vs. thymidylate synthase inhibitor (TS):

$$rs(S_{\text{DHFR}}, S_{\text{TS}}) = 772.25$$

$$n(S_{\text{DHFR}}, S_{\text{TS}}) = 216 \times 253 = 54,648$$

$$z = (772.3 - 12.241) / 6.349 \approx 119$$

Transformation of Z-scores to E-values. The background distribution of random-set raw scores of MDDR compounds conformed to an extreme value distribution (**Supplementary Fig. 1**, see also **Methods**). Given this observation, the probability of obtaining the same or better raw score by random chance alone can be calculated by:¹

$$P(Z > z) = 1 - \exp(-e^{-z\pi/\sqrt{6} - \Gamma'(1)})$$

¹ Pearson, W.R. Empirical statistical estimates for sequence similarity searches. *J Mol Biol* **276**, 71-84 (1998).

Where $\Gamma'(1)$ is the Euler-Mascheroni constant (≈ 0.577215665).

Then it follows that:

$$E(z) = P(z)N_{db}$$

Where N_{db} is the number of set comparisons made in the database search.

Note, however, that for Z-scores > 28 , $P(Z > z)$ exceeds the numerical precision of most computing languages (e.g., Python), and the following numerical approximation following a Taylor expansion can be used instead:²

$$x = -\exp(-z\pi / \sqrt{6}) - \Gamma'(1)$$

$$P(Z > z) = -(x + (x^2)/2 + (x^3)/6)$$

For the comparison of the sets Dihydrofolate reductase inhibitor (DHFR) vs. thymidylate synthase inhibitor (TS), $z \approx 119$, so:

$$P(Z > 119) \approx 2.92 \times 10^{-67}$$

$$E(119) = (2.92 \times 10^{-67})(246^2) \approx 1.77 \times 10^{-62}$$

The astute reader will note that this value does not perfectly match that reported in the **Results** for this comparison (1.11×10^{-61}) as the calculations in this example use fewer significant digits and the approximation error has accumulated.

² Valiant, P. *Personal communication*. 2004.