

## Supplementary information

Universal Prediction of Intramolecular H-bonds in organic structures P. T. A. Galek, L. Fábíán and F. H. Allen

### Summary of statistical assessment

Goodness-of-fit statistics: The fitted model equation is compared against an independent model (zero coefficients). The included statistics may be referred to in Galek et al. (2007).

Null hypothesis test: Determination of improvements in a fitted model over a null model (the 'null probability', in which  $P_0 = 0.29$  is always returned).

**Table S1.** Statistical assessment of generic IHB model fitting

#### S1A. Goodness-of-fit statistics

| Statistic                         | Independent | Full      |
|-----------------------------------|-------------|-----------|
| Observations                      | 21396       | 21396     |
| Sum of weights                    | 21396.000   | 21396.000 |
| DF                                | 21395       | 21378     |
| -2<br>Log(Likelihood)             | 27965.027   | 19335.796 |
| R <sup>2</sup> (McFadden)         | 0.000       | 0.309     |
| R <sup>2</sup> (Cox and<br>Snell) | 0.000       | 0.332     |
| R <sup>2</sup> (Nagelkerke)       | 0.000       | 0.455     |
| AIC                               | 27967.027   | 19371.796 |
| SBC                               | 27974.998   | 19515.273 |
| Iterations                        | 0           | 9         |

#### S1B. Test of the null hypothesis $H_0: Y=0.474$ (IHB exists)

| Statistic          | DF | $\chi^2$ | $Pr > \chi^2$ |
|--------------------|----|----------|---------------|
| -2 Log(Likelihood) | 17 | 8629.231 | < 0.0001      |
| Score              | 17 | 7363.540 | < 0.0001      |
| Wald               | 17 | 4869.499 | < 0.0001      |

**Table S2.** Statistical assessment of the *R5* IHB model fitting

#### S2A. Goodness-of-fit statistics

| Statistic                         | Independent | Full     |
|-----------------------------------|-------------|----------|
| Observations                      | 8184        | 8184     |
| Sum of weights                    | 8184.000    | 8184.000 |
| DF                                | 8183        | 8164     |
| -2<br>Log(Likelihood)             | 11324.074   | 9055.504 |
| R <sup>2</sup> (McFadden)         | 0.000       | 0.200    |
| R <sup>2</sup> (Cox and<br>Snell) | 0.000       | 0.242    |
| R <sup>2</sup> (Nagelkerke)       | 0.000       | 0.323    |
| AIC                               | 11326.074   | 9095.504 |
| SBC                               | 11333.084   | 9235.703 |
| Iterations                        | 0           | 9        |

S2B. Test of the null hypothesis  $H_0: Y=0.474$  (IHB exists)

| Statistic          | DF | $\chi^2$ | $Pr > \chi^2$ |
|--------------------|----|----------|---------------|
| -2 Log(Likelihood) | 19 | 2268.570 | < 0.0001      |
| Score              | 19 | 1809.238 | < 0.0001      |
| Wald               | 19 | 1271.616 | < 0.0001      |

**Table S3.** Statistical assessment of the *R6* IHB model fitting

S3A. Goodness-of-fit statistics

| Statistic                      | Independent | Full      |
|--------------------------------|-------------|-----------|
| Observations                   | 10291       | 10291     |
| Sum of weights                 | 10291.000   | 10291.000 |
| DF                             | 10290       | 10268     |
| -2 Log(Likelihood)             | 9569.416    | 5611.679  |
| R <sup>2</sup> (McFadden)      | 0.000       | 0.414     |
| R <sup>2</sup> (Cox and Snell) | 0.000       | 0.319     |
| R <sup>2</sup> (Nagelkerke)    | 0.000       | 0.527     |
| AIC                            | 9571.416    | 5657.679  |
| SBC                            | 9578.655    | 5824.176  |
| Iterations                     | 0           | 7         |

S3B. Test of the null hypothesis  $H_0: Y=0.826$  (IHB exists)

| Statistic          | DF | $\chi^2$ | $Pr > \chi^2$ |
|--------------------|----|----------|---------------|
| -2 Log(Likelihood) | 22 | 3957.737 | < 0.0001      |
| Score              | 22 | 3863.099 | < 0.0001      |
| Wald               | 22 | 1726.761 | < 0.0001      |

**Table S4.** Statistical assessment of the *R7* IHB model fitting

S4A. Goodness-of-fit statistics

| Statistic                      | Independent | Full     |
|--------------------------------|-------------|----------|
| Observations                   | 2923        | 2923     |
| Sum of weights                 | 2923.000    | 2923.000 |
| DF                             | 2922        | 2906     |
| -2 Log(Likelihood)             | 4029.159    | 2987.888 |
| R <sup>2</sup> (McFadden)      | 0.000       | 0.258    |
| R <sup>2</sup> (Cox and Snell) | 0.000       | 0.300    |
| R <sup>2</sup> (Nagelkerke)    | 0.000       | 0.401    |
| AIC                            | 4031.159    | 3021.888 |
| SBC                            | 4037.139    | 3123.554 |
| Iterations                     | 0           | 6        |

S4B. Test of the null hypothesis  $H_0: Y=0.456$  (IHB exists)

| Statistic          | DF | $\chi^2$ | $Pr > \chi^2$ |
|--------------------|----|----------|---------------|
| -2 Log(Likelihood) | 16 | 1041.271 | < 0.0001      |
| Score              | 16 | 916.434  | < 0.0001      |
| Wald               | 16 | 688.025  | < 0.0001      |

**Table S5.** Classification table of generic IHB model training observations

| observed\<br>predicted | <i>False</i> | <i>True</i> | Total | % correct |
|------------------------|--------------|-------------|-------|-----------|
| <i>False</i>           | 4927         | 2779        | 7706  | 63.94%    |
| <i>True</i>            | 1982         | 11708       | 13690 | 85.52%    |
| Total                  | 6909         | 14487       | 21396 | 77.75%    |

**Table S6.** Classification table of *R5* IHB model training observations

| observed\<br>predicted | <i>False</i> | <i>True</i> | Total | % correct |
|------------------------|--------------|-------------|-------|-----------|
| <i>False</i>           | 3118         | 1183        | 4301  | 72.49%    |
| <i>True</i>            | 1218         | 2665        | 3883  | 68.63%    |
| Total                  | 4336         | 3848        | 8184  | 70.66%    |

**Table S7.** Classification table of *R6* IHB model training observations

| observed\<br>predicted | <i>False</i> | <i>True</i> | Total | % correct |
|------------------------|--------------|-------------|-------|-----------|
| <i>False</i>           | 1489         | 320         | 1809  | 82.31%    |
| <i>True</i>            | 1582         | 6900        | 8482  | 81.35%    |
| Total                  | 3071         | 7220        | 10291 | 81.52%    |

**Table S8.** Classification table of *R7* IHB model training observations

| observed\<br>predicted | <i>False</i> | <i>True</i> | Total | % correct |
|------------------------|--------------|-------------|-------|-----------|
| <i>False</i>           | 1112         | 479         | 1591  | 69.89%    |
| <i>True</i>            | 268          | 1064        | 1332  | 79.88%    |
| Total                  | 1380         | 1543        | 2923  | 74.44%    |

**Fig. S1.** Distributions of normalised donor-acceptor distance  $r_{norm}$  in (a) *R5*, (b) *R6* & (c) *R7* IHBs in the CSD (see manuscript for definition of  $r_{norm}$ ).

