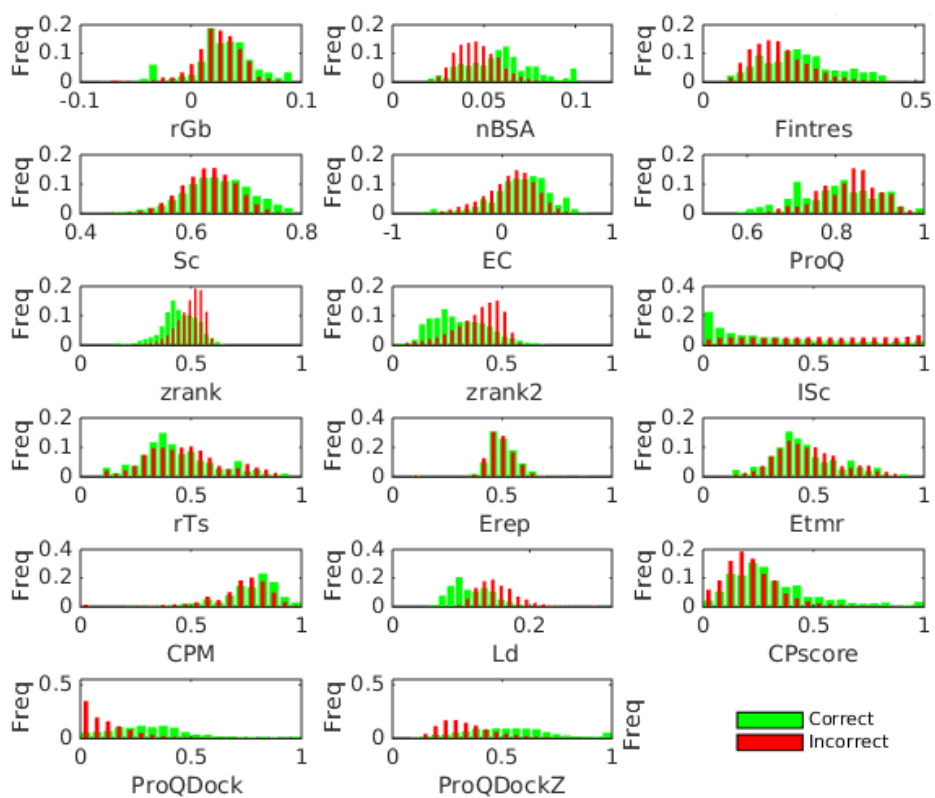**Fig. S1.** Distribution of training features on the MOAL set for correct (green) and incorrect (red) models. The features that are outside the -1,1 range are scaled using logistic scaling (see Methods) to put them in the range 0-1. For the methods that are energy terms, zrank, zrank2, Isc, rTs, Erep, and Etmr lower scores are better, for the others higher are better.
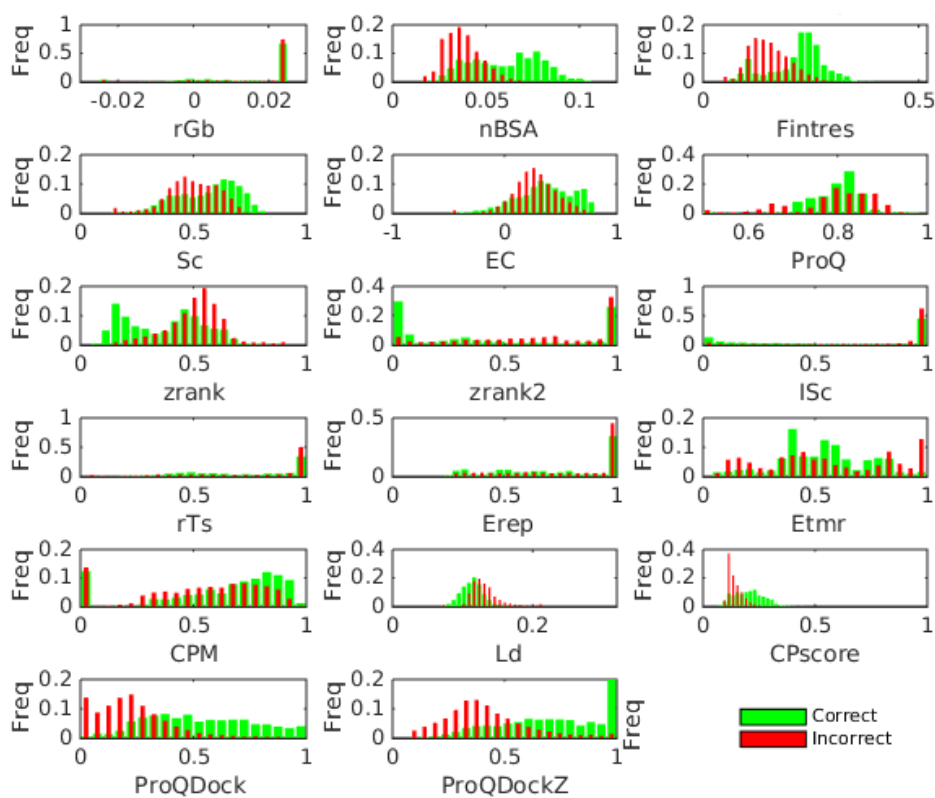
Fig. S2. Distribution of training features on the CAPRI set for correct (green) and incorrect (red) models. The features that are outside the -1,1 range are scaled using logistic scaling (see Methods) to put them in the range 0-1. For the methods that are energy terms, zrank, zrank2, Isc, rTs, Erep, and Etmr lower scores are better, for the others higher are better.