Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2018

Supplementary Information for Physical Chemistry Chemical Physics
A unified scoring function for protein folding and drug-binding pocket recognition.

Anna Battisti[1], Stefano Zamuner[2], Edoardo Sarti[3], and Alessandro Laio[1]
[1] *International School for Advanced Studies (SISSA), Via Bonomea 265, I-34136 Trieste, Italy*
[2] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Institute of Physics Laboratory of Statistical Biophysics 1015 Lausanne Switzerland*
[3] *National Institute of Neurological Disorders and Stroke, NIH, 35 Convent Dr. Bethesda MD 20894, USA.*

SASA parameters optimization

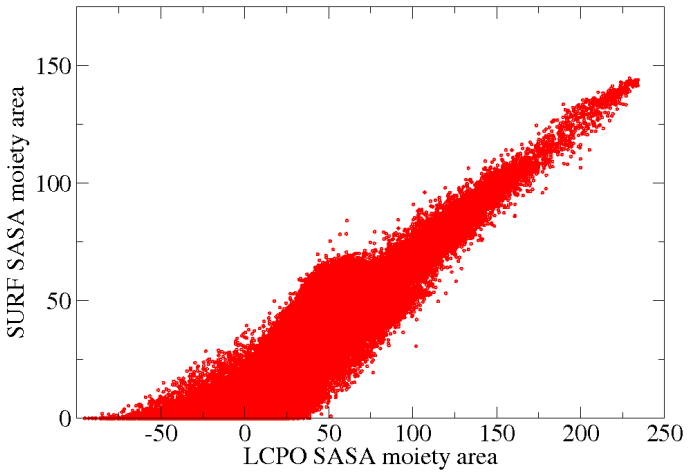


FIG. S1: Correlation between the SASA of the moieties, obtained with the MLCPO algorithm (see the main text) and perfomed by the SURF tool of VMD. The atomic radius r_i and the $P1 - P4$ MLCPO parameters given in Tab.I of the main text have been varied to have a resonable Pearson correlation value (0.89).

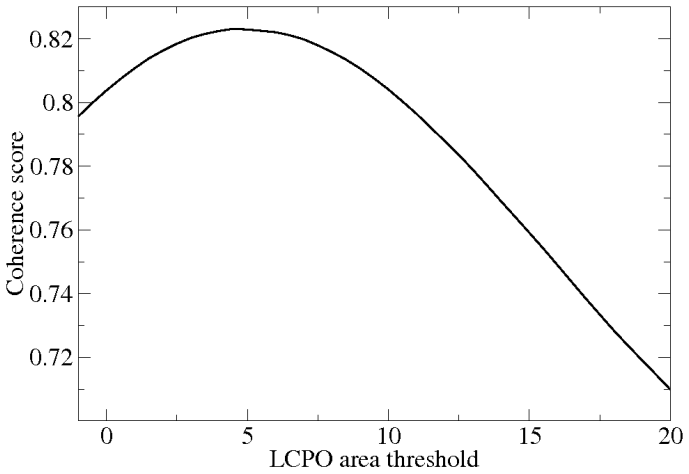


FIG. S2: The coherence score defined as the fraction of moieties in a structure for which SURF and the MLCPO algorithm agree on the environmental class assignation (buried or exposed) as a function of the solvent exposure classification threshold on MLCPO. The SURF threshold is fixed at 1.5 Å. The maximum value (82%) is obtained at the threshold value $A_t = 5$ Å^2.

Cross Validation

A		B		C	
Training	Validation	Training	Validation	Training	Validation
T0451 T0468	T0388	T0388 T0397	T0451	T0388 T0397	T0541
T0472 T0485	T0397	T0415 T0425	T0468	T0415 T0425	T0544
T0488 T0504	T0415	T0427 T0432	T0472	T0427 T0432	T0562
T0511 T0517	T0425	T0433 T0437	T0485	T0433 T0437	T0563
T0522 T0526	T0427	T0440 T0445	T0488	T0440 T0445	T0569
T0539 T0541	T0432	T0447 T0541	T0504	T0447 T0451	T0575
T0544 T0562	T0433	T0544 T0562	T0511	T0468 T0472	T0579
T0563 T0569	T0437	T0563 T0569	T0517	T0485 T0488	T0594
T0575 T0579	T0440	T0575 T0579	T0522	T0504 T0511	T0623
T0594 T0623	T0445	T0594 T0623	T0526	T0517 T0522	T0626
T0626 T0628	T0447	T0626 T0628	T0539	T0526 T0539	T0628

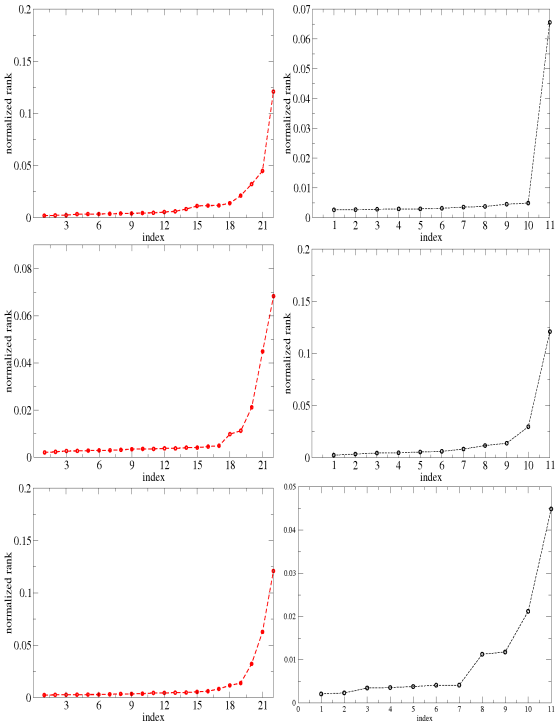


FIG. S3: (Left) In the Table we show how we partitioned in complementary subsets the CASP8/9 decoy sets in order to make the cross-validation. The partition in A, B, C allow including each decoy in a validation set. The parameters r_1, r_2, r_3 and r_4 entering in the definition of the class (see main text) are optimized on the 22 training decoy sets, indicated in the Table, by employing the L-BFGS-B algorithm and minimizing the sum of the normalized native rank; (Right, Left panel): the optimized ranks sorted by increasing value. (Right, Right panel): the ranks on the validation sets, sorted by increasing values.

The BACH-MOI class definition

In order to allow for a smooth change between the class around the distance thresholds we introduce a smoothing function $f(r_{ij})$ estimated for the distance r_{ij} between moiety i and moiety j.

$$f_\alpha(r_{ij}) = \frac{1-(\frac{r_{ij}}{r_\alpha})^n}{1-(\frac{r_{ij}}{r_\alpha})^{2n}} \quad , \quad \alpha \in \{1,2,3\} \tag{S1}$$

Here, r_α represents one of the three distance thresholds r_1, r_2, r_3 for a moiety pair (r_4 is excluded, as there is no scoring class for $r > r_4$). We fixed the parameter n to 30 for r_1 and 50 for r_2 and r_3. For each r_α threshold we calculated the two roots $r_\alpha^{\min}$ and $r_\alpha^{\max}$ for which $f_\alpha = 0.95$ and $1 - f_\alpha = 0.95$ respectively:

$$r_\alpha^{\min} = r_\alpha * \left(\frac{1-0.95}{0.95}\right)^{(1/n)} \quad , \quad r_\alpha^{\max} = r_\alpha * \left(\frac{0.95}{1-0.95}\right)^{(1/n)}$$

If $r_{ij} < r_\alpha^{\min}$ the pair of moieties is in class α; if $r_{ij} > r_\alpha^{max}$ it is in class $\alpha+1$; if $r_\alpha^{min} < r_{ij} < r_\alpha^{max}$ it is in classes α and $\alpha+1$ respectively with a weight $f_\alpha(r_{ij})$ and $1 - f_\alpha(r_{ij})$. Classes α and $\alpha+1$ have the same SASA exposure (either b/b, e/b or e/e), unless $\alpha+1$ is the non-contact class, in which case each of the three classes in the range $r_2 < r \leq r_3$ mixes with the only class in the range $r_3 < r \leq r_4$.

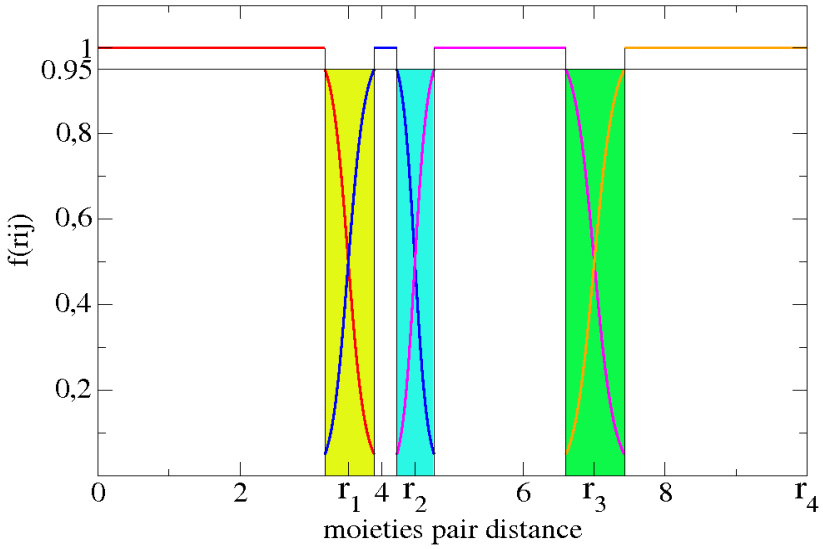


FIG. S4: The smooth function gives the weight for the class assignment of a moiety pair. If r_{ij} is in the region represented with the colored box centered on r_α, the moiety pair belong to the adjacent classes with a weight given by $f_\alpha(r_{ij})$ and $1 - f_\alpha(r_{ij})$, while outside the colored boxes the class is assigned with probability 1.

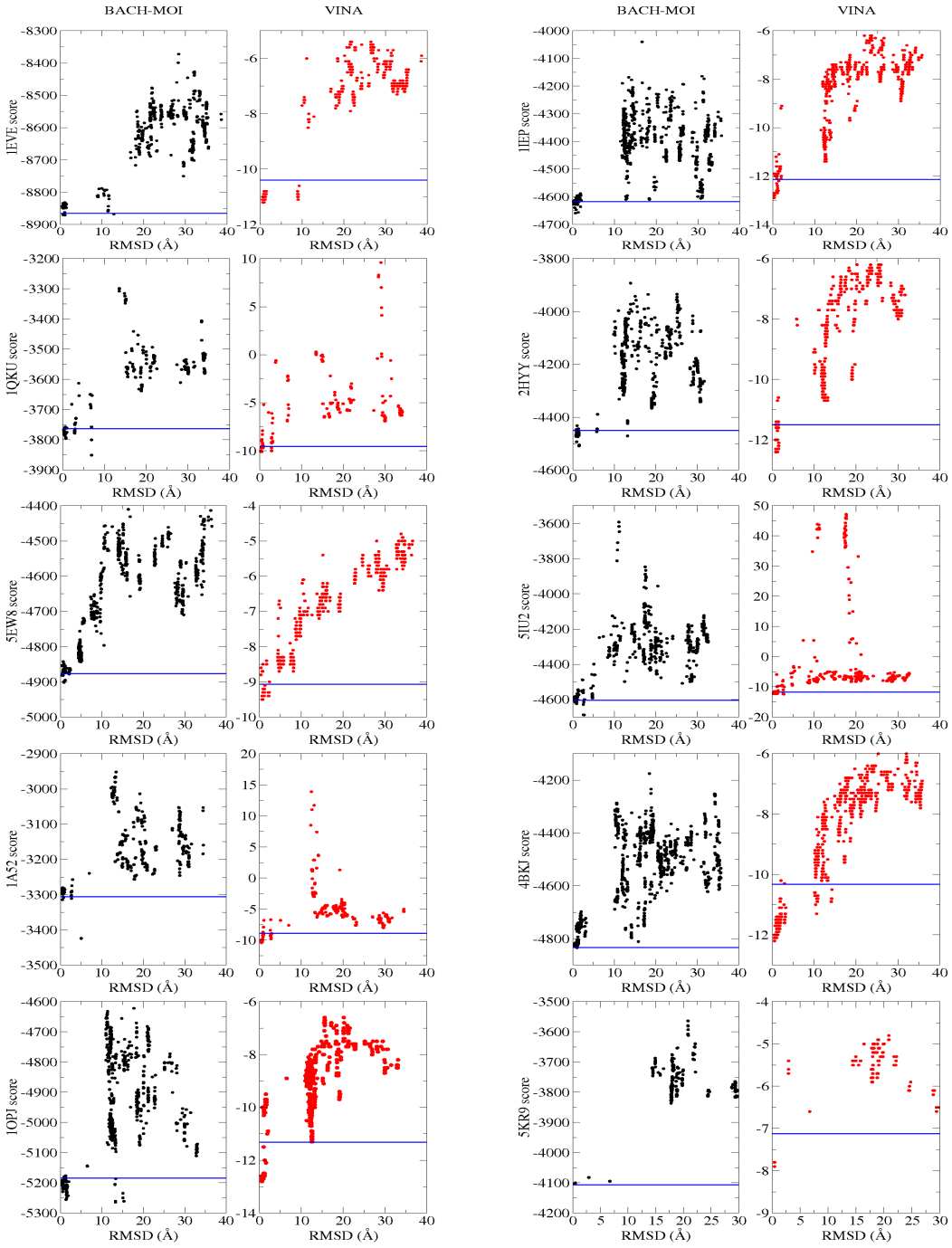


FIG. S5: The score and RMSD correlation on 10 drug/protein complexes, the PDB entry of which is given in the label. For each complex the analysis was made with the BACH-MOI scoring function (black circle) and with VINA scoring function (red circle). The blue line highlights the score of the experimental structure.

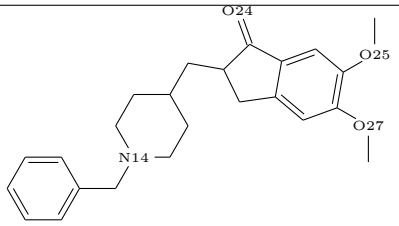
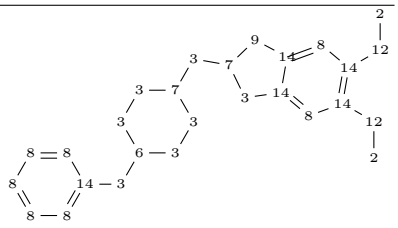
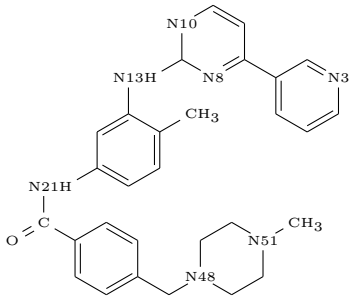
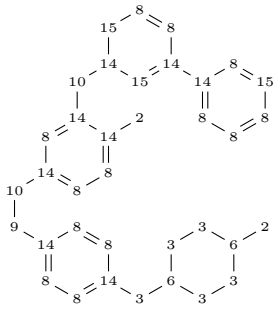
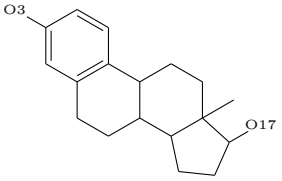
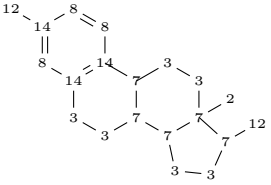
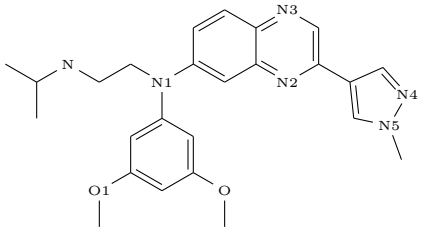
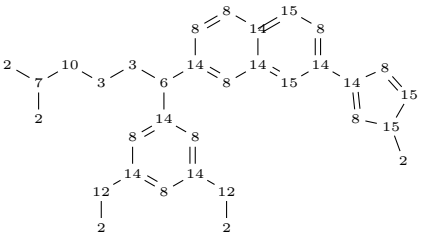
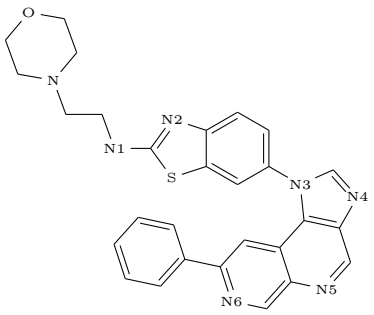
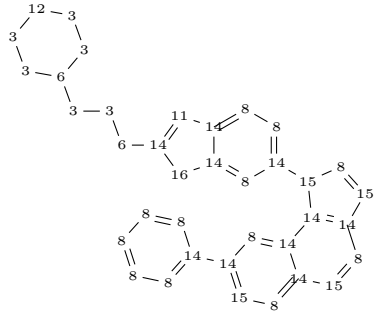
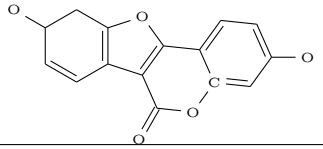
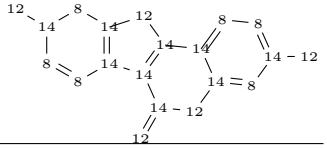
Ligand	PDB entry	2D Diagram	2D Diagram in moieties representation
ARICEPT ($C_{24}H_{29}NO_3$)	1EVE		
IMATINIB ($C_{29}H_{31}N_7O$)	1IEP 2HYY 4BKJ 1OPJ		
ESTRADIOL ($C_{18}H_{24}O_2$)	1QKU 1A52		
JNJ-4275693 ($C_{25}H_{30}N_6O_2$)	5EW8		
6DA ($C_{28}H_{25}N_7OS$)	5IU2		
COUMESTROL ($C_{15}H_8O_5$)	5KR9		

FIG. S6: For each protein complex identified by the PDB code we report here the name and chemical formula of the ligand, as well as its 2D diagram in standard representation and in the moiety code representation (see Fig. 7 in the main text.)