Electronic supplementary materials *Mendeleev Commun.*, 2023, **33**, 802–805

On importance of explicit account of non-complementary contacts in scoring functions

Arslan R. Shaimardanov, Dmitry A. Shulga and Vladimir A. Palyulin

Materials and methods

In order to quantify both complementary and non-complementary ligand receptor contacts the concept of solvent accessible surface area (SASA) was employed.

CSA

We introduce Contact Surface Area (CSA) as a measure to estimate density of intermolecular contacts between protein and ligand. Hereinafter *CSA[lig_type+prot_type]* denotes contact area between ligand atoms which have *lig_type* atomic type and protein atoms which have *prot_type* atomic type. CSA was calculated as follows. First, SASA's of free (non-bonded) ligand and protein were measured and per-atom SASA values were obtained. Second, atoms of the corresponding types, which have non-zero SASA, were selected. Third, the "complex" was created of those selected atoms and change/loss of solvent accessible area (ΔSASA) was measured. This change was considered a value of the area of newly formed contacts between different types of ligand and protein atoms, i.e., CSA. The source code is available at http://molmodel.com/hg/dSAS/ .

CSA-based scoring function

Several models were tested in this work. All of them are based on a simple linear regression, including different CSA terms (S1). Each model was adjusted to reproduce the reference (experimental) free binding energy values, provided in the CASF-2016 database. [10.1021/acs.jcim.8b00545] The models were trained and tested on the same set of molecules. The models were not thoroughly validated as the work was targeted on highlighting deficiencies of the currently existing models and not on proposing a new one which would be ready to use.

$$\Delta G = -RT \cdot lnK_{i/d} = b + \sum_i w_i \times CSA_i \quad \text{(S1)}$$

The first model was based purely on complementary types of contacts, meaning CSAi∈{CSA[HYD+HYD], CSA[DON+ACC], CSA[ACC+DON]} (Figure S1).

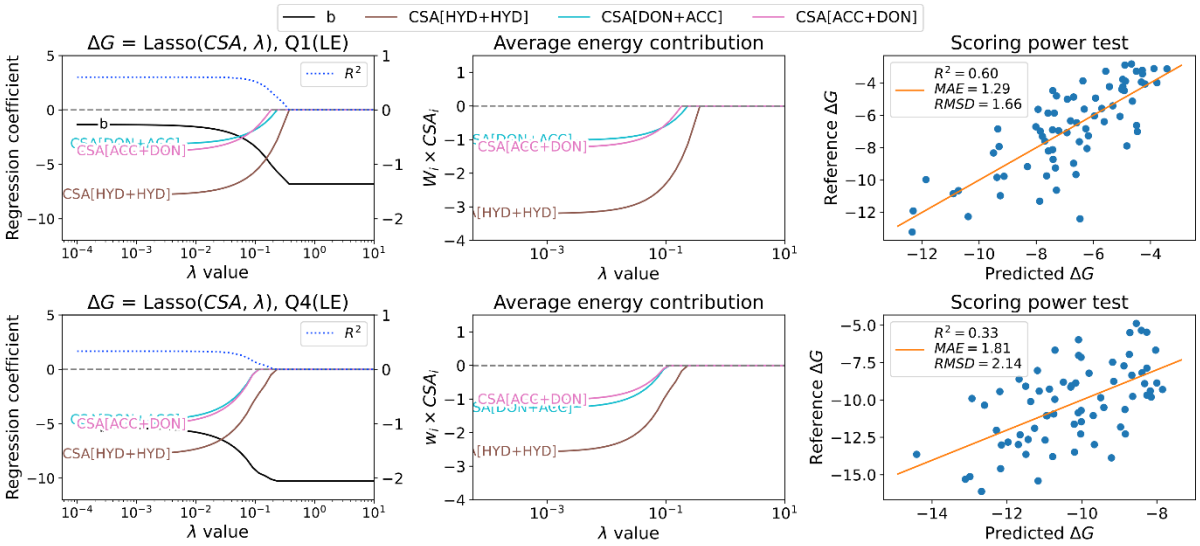

Figure S1 Lasso regression coefficients and total energy contribution of complementary contacts for Q1 and Q4 complexes.

The second one also included non-complementary types of contacts, meaning CSAi∈{CSA[HYD+HYD], CSA[DON+ACC], CSA[ACC+DON], CSA[HYD+DON], CSA[DON+HYD], CSA[HYD+ACC], CSA[ACC+HYD]} (Figure S2).

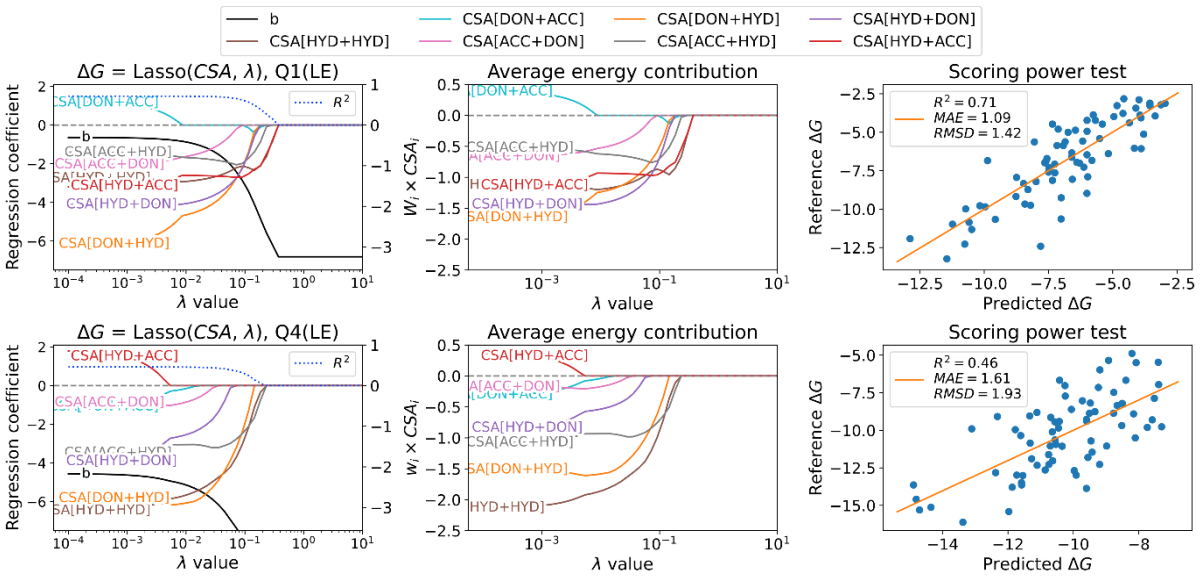

Figure S2 Lasso regression coefficients and total energy contribution of complementary and non-complementary contacts for Q1 and Q4 complexes.

The last model also included contacts, involving atoms of unspecified type (denoted by *OTH*). It served as a model with the highest achievable quality in the current settings. In this model. the following list of descriptors was included: $CSA_i \in \{CSA[\mathrm{HYD+HYD}], CSA[\mathrm{HYD+DON}], CSA[\mathrm{HYD+ACC}], CSA[\mathrm{HYD+OTH}], CSA[\mathrm{DON+HYD}], CSA[\mathrm{DON+DON}], CSA[\mathrm{DON+ACC}], CSA[\mathrm{DON+OTH}], CSA[\mathrm{ACC+HYD}], CSA[\mathrm{ACC+DON}], CSA[\mathrm{ACC+ACC}], CSA[\mathrm{ACC+OTH}], CSA[\mathrm{OTH+HYD}], CSA[\mathrm{OTH+DON}], CSA[\mathrm{OTH+ACC}], CSA[\mathrm{OTH+OTH}]\}$.

Tools

SASA and CSA values were calculated using PyMOL python API. [https://pymolwiki.org/index.php/Get_area] Solvent radius was set to 1.4A and dots density was set to 3.

CASF-2016 core set was used as a source of high-quality 3D structures and experimental $pK_{i/d}$ values for protein-ligand complexes. [10.1021/acs.jcim.8b00545]

OpenBabel python API was used to assign atomic types (DON, ACC and HYD) using predefined SMARTS patterns (Table S1).

Table S1 SMARTS patterns for atomic types.

SMARTS pattern	Description	Type
[*]	Any atom	OTH
[#15,#16]	Common Sulphur and Phosphorous	
[#8,#16;R]	Heterocyclic Oxygen and Sulphur	
[#8,#16;H1]	Oxygen/Sulphur with 1 attached hydrogen	DON
[#7;H1,H2,H3]	Amine and amide nitrogen	
[#7X3;H1,H2,H3]		
[#7;!H0;+0,+1]	Amine or ammonia nitrogen	
[#7X3H2]~[#6]~[#7X2H1]	Amidine nitrogen	
[#7X2H1]~[#6]~[#7X3H2]		
[#7X3H2]~[#6]~[#7X3H2]	Amidine nitrogen [+1]	
[#7X3;H1,H2,H3]	Amine nitrogen	
[#8,#16;X1H0]~[#6]	O=C or S=C	ACC
[#8X2H1]~[#6]~[#8X1H0]	Carboxyl and ester oxygen	
[#8X1H0]~[#6]~[#8X1H0]		
[#8X1]~[#6]~[#8X1H0]		
[#8X1]~[#15,#16]	Phospho-/Sulfo-group Oxygen	
[#7X2H0]	Heterocyclic nitrogen	
[#6X1](~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])	Any carbon with no polar neighbors	HYD
[#6X2](~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])(~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])		
[#6X3](~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])(~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])(~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])		
[#6X4](~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])(~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])(~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])(~[!$([#7,#8,#15,#16,#9,#17,#35,#53])])		

Table S2 Statistical metrics of different linear regression models aimed to reproduce free binding energy.

	R^2 (scoring power)		MAE, kcal/mol		RMSD, kcal/mol	
	Q1	**Q4**	**Q1**	**Q4**	**Q1**	**Q4**
Complementary only	0.60	0.33	1.29	1.81	1.66	2.14
Complementary + non-complementary	0.71	0.46	1.09	1.61	1.42	1.93
Complementary + non-complementary + CSA[DON-DON] + CSA[ACC-ACC]	0.72	0.46	1.08	1.62	1.40	1.92
Complementary + non-complementary CSA[DON-DON] + CSA[ACC-ACC] + CSA[OTH-*] + CSA[*-OTH]	0.76	**0.71**	1.00	1.19	1.29	1.42

Table S3 Absolute value of Pearson correlation coefficient between CSA of different types of contacts.

	CSA[ACC+ACC]	CSA[ACC+DON]	CSA[ACC+HYD]	CSA[DON+ACC]	CSA[DON+DON]	CSA[DON+HYD]	CSA[HYD+ACC]	CSA[HYD+DON]	CSA[HYD+HYD]
CSA[ACC+ACC]	1.00	0.75	0.61	0.24	0.23	0.13	0.24	0.15	0.02
CSA[ACC+DON]	0.75	1.00	0.63	0.05	0.20	0.02	0.17	0.26	0.00
CSA[ACC+HYD]	0.61	0.63	1.00	0.02	0.01	0.20	0.13	0.07	0.22
CSA[DON+ACC]	0.24	0.05	0.02	1.00	0.85	0.71	0.19	0.04	0.20
CSA[DON+DON]	0.23	0.20	0.01	0.85	1.00	0.68	0.08	0.13	0.26
CSA[DON+HYD]	0.13	0.02	0.20	0.71	0.68	1.00	0.04	0.12	0.08
CSA[HYD+ACC]	0.24	0.17	0.13	0.19	0.08	0.04	1.00	0.77	0.61
CSA[HYD+DON]	0.15	0.26	0.07	0.04	0.13	0.12	0.77	1.00	0.55
CSA[HYD+HYD]	0.02	0.00	0.22	0.20	0.26	0.08	0.61	0.55	1.00