Available online at www.sciencedirect.com

ScienceDirect

Mendeleev Commun., 2023, **33**, 802–805

Mendeleev Communications

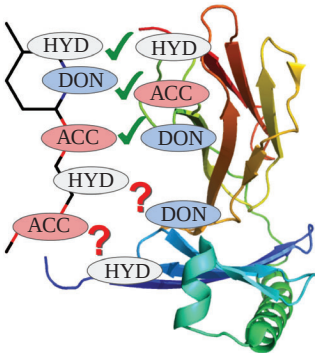
On importance of explicit account of non-complementary contacts in scoring functions

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

Department of Chemistry, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation. E-mail: vap@qsar.chem.msu.ru

DOI: 10.1016/j.mencom.2023.10.021

Based both on the practice of post-processing by a human expert and on the higher values of the accuracy metrics of machine learning scoring functions, it is suggested that when estimating the free energy of binding in a ligand–receptor complex, a significant part of intermolecular interactions is still not explicitly taken into account. An assessment is made of how explicit consideration of non-complementary ligand–receptor interactions could improve the accuracy of the description of contemporary classical scoring functions, which tend to use only terms of complementary/favorable interactions.



Keywords: non-complementary contacts, unfavorable contacts, binding free energy, scoring functions, molecular modeling, molecular docking, protein–ligand interactions.

Current drug discovery is a very complex and risky process, involving multifactor optimization and numerous cross-functional studies. Computer-aided drug discovery helps simplify the discovery process and reduce some of the risks.[1,2] Once the structure of a drug target is well defined, it is often used in conjunction with a set of structure-based drug discovery tools. Perhaps the most widely used in hit finding and hit-to-lead processes is virtual screening using molecular docking, in which the most plausible binding modes, as well as free energy estimations, are performed *in silico* on large libraries of potential ligands. Fast estimation of the free energy of ligand–receptor affinity during docking is carried out using scoring functions (SFs). Currently, four types of SFs have been proposed, namely physics (force field) based SFs, empirical SFs, knowledge based SFs and machine learning based SFs.[3,4]

SFs are ubiquitously and successfully used in numerous drug development projects.[5] By design, SFs are a very rough approximation of the free energy of binding of a ligand to a receptor at a particular ligand–receptor position, their main advantage is that SFs enable structure-based virtual screening, used in the early stages of drug discovery to narrow down the number of structures studied subsequently. The downside of the roughness and high throughput of SFs is the need for post-processing of the initial results using other techniques[6] and the involvement of expert visual inspection.[7] The fact that a human expert can rule out unnatural and unlikely binding modes suggests that there must be quite significant criteria for the quality of a binding mode that are currently still beyond the consideration of contemporary classical SFs. Additionally, a recent class of machine learning SFs have been able to overcome the accuracy limit previously achieved by classical SFs,[8,9] although there is still reasonable doubt that at least part of the success may be due to overfitting.[10–13] We hypothesize that one of the strong assumptions in the SFs is that currently only favorable interactions are generally considered. In this regard, given the multiple possible interactions, human expert visual inspection and machine learning SFs can go beyond the 'favorable interactions only' pattern. To the best of our knowledge, there has been no systematic study of the usefulness of including unfavorable or non-complementary interactions. The only exception is the work describing the parameterization of such SF as ScorpionScore,[14] in which mismatched interactions were explicitly taken into account, but with little discussion and generalization.

The most common types of specific intermolecular interactions[15] considered in SFs are hydrophobic (HYD) interactions and hydrogen bonding, which are believed to constitute the major part of drug-like ligand–receptor interactions.[16] Naturally, HYD ligand–HYD receptor interactions favorably influence the score, as do hydrogen bonding interactions in complementary donor (DON) ligand–acceptor (ACC) receptor and ACC ligand–DON receptor pairs. Thus, complementary ligand–receptor interactions include those designated as Lig(HYD)–Rec(HYD), Lig(ACC)–Rec(DON) and Lig(DON)–Rec(ACC).

This work aims to explore the extent to which explicit consideration of non-complementary ligand–receptor interactions may play a role in better characterizing ligand–receptor interactions. Therefore, we specifically investigated how explicitly accounting for non-complementary interactions, such as Lig(HYD)–Rec(DON), Lig(HYD)–Rec(ACC), Lig(ACC)–Rec(HYD), Lig(DON)–Rec(HYD), Lig(DON)–Rec(DON) and Lig(ACC)–Rec(ACC), helps to describe the experimental free energy of ligand–receptor interactions.

To this end, CASF-2016 Update,[17] a well-established benchmark for evaluating SFs and the core set of the PDBBind database,[18,19] is used as a source of reference free energy. The complementary and non-complementary contact surfaces, called Contact Surface Areas (CSAs), are used as descriptors to build linear regularized models (LASSO[20]) in order to qualitatively and quantitatively

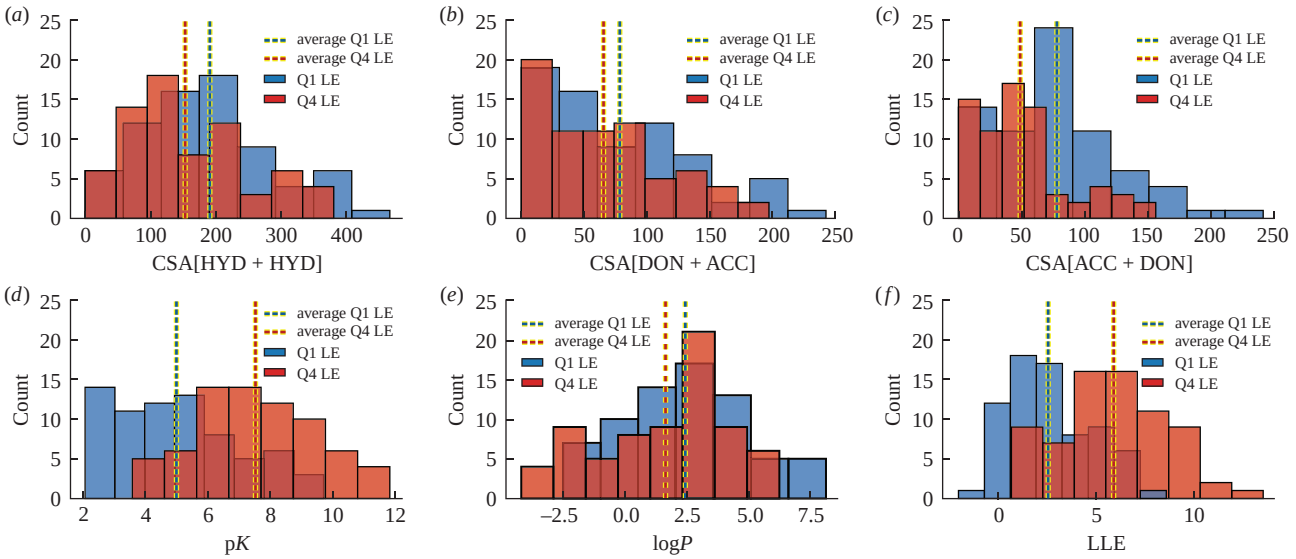


Figure 1 Distribution of the main characteristics of complexes with ligands having the lowest (Q1) and highest (Q4) LE values: (*a*) CSA[HYD+HYD], (*b*) CSA[DON+ACC], (*c*) CSA[ACC+DON], (*d*) $pK_{i/d}$, (*e*) log*P* and (*f*) LLE.

interpret the obtained results. LASSO models include a regularization coefficient λ, larger values of which lead to zeroing out the contribution of less significant descriptors to the model.

Hereinafter, 'CSA[lig_type+prot_type]' denotes the contact area between ligand atoms, which have the 'lig_type' atomic type, and protein atoms, which have the 'prot_type' atomic type.

At the first stage of the research, we analyzed the distribution of complementary and non-complementary contact areas of the ligands.

The CASF-2016 core set was chosen as a standard and reliable source of both experimental 3D structures and binding affinities.[17] Because of its certain bias towards generally well-known and favorable complexes, the number of non-complementary contacts is *a priori* expected to be low, so we decided to test this hypothesis explicitly. The complexes were divided into quartiles according to the ligand efficiency (LE) value, which is a measure of the binding affinity per heavy (non-hydrogen) atom. It was calculated as the negative decimal logarithm of the inhibition constant (pK_i) divided by the number of heavy atoms. The first quartile (Q1) contained 72 ligands with low 25% LE values, and the fourth quartile (Q4) contained 71 ligands with high 25% LE values, which means that Q4 ligands are quite more optimal than Q1 ones.

The idea of dividing a set of molecules was to outline the dissimilarities between its parts in a simple and intuitive way. In this particular case, it was assumed that the low LE (and hence lower density per atom, $pK_{i/d}$) of ligands in Q1 is a direct consequence of the much more frequent presence of non-optimal (*i.e.*, non-complementary) contacts compared to highly optimized ligands with high LE values in Q4. Thus, by comparing the distinct characteristics of the two subsets, we could see more clearly the importance of the explicit account of such contacts within the SF.

We first compared the CSA values for the complementary contacts (Figure 1).

The average CSA of the complementary contacts for Q1 ligands is higher than for Q4 ligands, but the average binding affinity ($pK_{i/d}$) is much higher for Q4 ligands. It seems that Q1 ligands are generally more lipophilic, potentially hampering their ADMET properties.[21] Accordingly, lipophilic ligand efficiency (LLE),[21] balancing affinity and lipophilicity, is also higher for Q4 ligands in general and above the recommended threshold value of 5–6.[21] It can be assumed that the higher the complementary

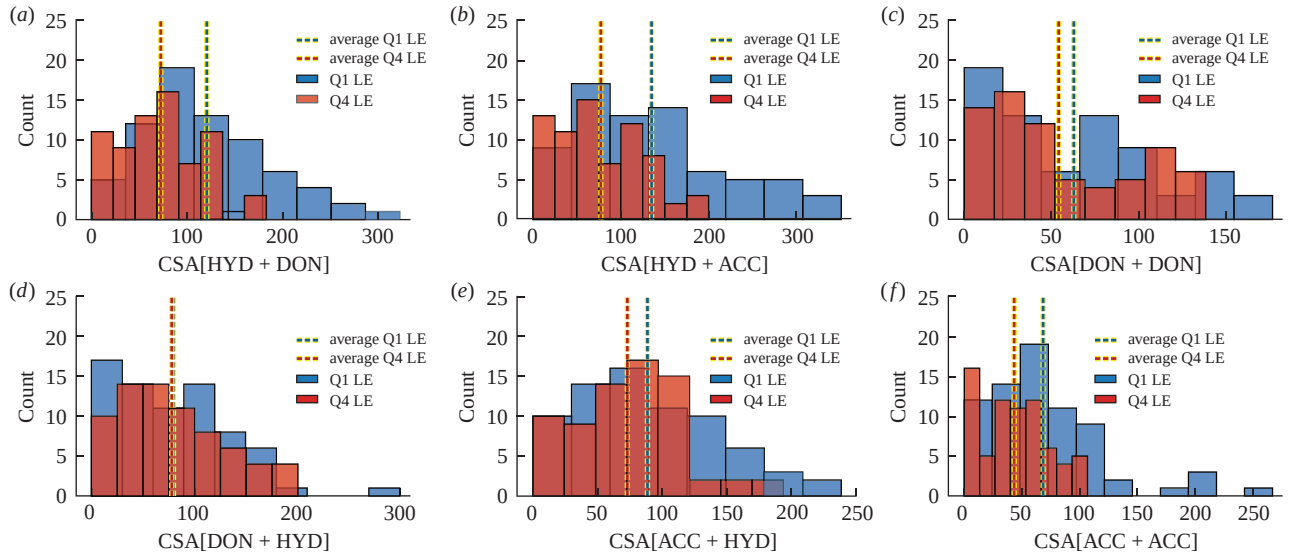


Figure 2 Distribution of CSA values of the non-complementary contacts in complexes with ligands having the lowest (Q1) and highest (Q4) LE values: (*a*) CSA[HYD+DON], (*b*) CSA[HYD+ACC], (*c*) CSA[DON+DON], (*d*) CSA[DON+HYD], (*e*) CSA[ACC+HYD] and (*f*) CSA[ACC+ACC].

CSA, the lower the binding constant (*i.e.*, the weaker the binding), which is contrary to common sense and the practice of chemists.[7] This suggests that another source of correction, presumably a non-complementary CSA (Figure 2), is needed.

However, the CSA of non-complementary contacts is also larger on average for Q1 ligands with low LE. It is assumed that such contacts are energetically unfavorable and must be compensated for by favorable contacts, thereby degrading the resulting binding energy. This means that they should be taken into account in the model describing binding energy.

At the second stage, we test the hypothesis about the importance of explicitly taking into account non-complementary contacts. To this end, we build several regression models that reproduce the reference experimentally measured binding free energy.

The first model only considers complementary contacts, such as CSA[HYD+HYD], CSA[DON+ACC] and CSA[ACC+DON]. The model was tested separately for Q1 and Q4 ligands. All descriptor values were normalized to range from 0 to 1, so b in equation (1) can be viewed as the minimum possible predicted value of ΔG on the set. At the same time, the value of b at high regularization coefficients (*i.e.*, when $w_i = 0$) can be considered as the average ΔG on the set.

$$\Delta G_{bind} = -RT \ln K_{i/d} = b + \sum_i w_i \mathrm{CSA}_i$$
$$\mathrm{CSA}_i \in \{\mathrm{CSA[HYD + HYD]}, \mathrm{CSA[DON + ACC]}, \mathrm{CSA[ACC + DON]}\}, \quad (1)$$

where ΔG_{bind} is the binding free energy, R is the molar gas constant, T is the absolute temperature, $K_{i/d}$ is the inhibition/dissociation constant, CSA_i is the CSA of the specific type, w_i is the fitted regression coefficient, and b is the fitted intercept value.

First, it can be seen (Figure S1, see Online Supplementary Materials) that HYD interactions make the largest contribution in terms of binding affinity (energy). In addition, Lig(DON)–Rec(ACC) and Lig(ACC)–Rec(DON) interactions have the same statistical weight and make a comparable contribution to the binding free energy value. Both types of interactions are considered favorable in the regression model, with a significantly larger contribution from HYD interactions compared to hydrogen bonding interactions. The latter is explained by both enthalpy and entropy contributions.[15] Overall, these observations are consistent with generally accepted knowledge in the field of medicinal chemistry,[15] which confirms the validity of the model.

Second, it should be noted that the R^2 value in the scoring test for both ligands Q1 and Q4 corresponds to the range of R^2 values obtained in the CASF-2016 benchmark[17] for a diverse set of practically used SFs, which further confirms the validity and reasonableness of CSA as means of estimating the interaction measure.

Third, there is a significant difference between ligands Q1 and Q4. The baseline b (minimum and average predicted ΔG values) is significantly larger for Q4 ligands. The total contribution of HYD contacts is slightly less for Q4 ligands, while DON–ACC contacts are at the same level for both Q1 and Q4. In addition, the prediction quality (R^2, MAE and RMSD) for Q4 complexes is drastically worse, meaning that there is a significant difference between the binding modes in these quartiles. One might speculate that the current model based only on complementary contacts is not sufficient to capture all kinds of valuable interactions involved in binding, and non-complementary contacts could be included as a correction. However, this discrepancy between predicted and experimental energies may also be related to the unavoidable uncertainties in the experimental data (*e.g.*, imprecise 3D geometries and chemical structures, missing water-mediated interactions and entropy contribution), as well as due to the crude SF approximation that associates a single protein–ligand conformation to an experimental value of binding affinity, which in practice reflects the affinity of the conformational ensemble. Next, we test only the first assumption about the importance of non-complementary contacts.

In the second model, in addition to complementary contacts, non-complementary polar–HYD contacts are taken into account. DON–DON and ACC–ACC contacts were excluded from the model due to their purely electrostatic nature, in contrast to polar–HYD contacts.

Compared to the model that only considers complementary contacts (see Figure S1), the model that additionally considers non-complementary contacts (Figure S2) performs numerically better (Table 1) in terms of R^2, MAE and RMSD for both Q1 and Q4 complexes. However, the reduction in residual error (both MAE and RMSD) is an order of magnitude lower than the residual error itself.

It is also clear from Figure S2 that the quality of the extended model for Q4 ligands is still noticeably worse than the quality of the model for Q1, which means that there are certain energetic terms that have yet to be taken into account but are not observed in the current model. An additional experiment taking into account all types of atoms not included in the previous definitions (denoted by the atomic type 'OTH') unexpectedly leads to a significant improvement in the statistics of the produced linear models (Table S3, see Online Supplementary Materials), especially for the Q4 part of the set ($R^2 = 0.71$, MAE = 1.19 and RMSD = 1.42). The latter means that seemingly important interactions, and possibly their cooperativity, are currently not accounted for in contemporary classical SFs, and also provides an additional explanation for the relative success of the machine learning SFs. However, it should be noted that inclusion of all ligand and receptor atom type pairs as well as 'OTH' effectively results in doubling the number of CSA descriptors.

However, this minor numerical improvement comes at the cost of a loss of qualitative interpretability (and potentially practical applicability). For example, the contacts Lig(DON)–Rec(ACC) are considered unfavorable in this model (see Figures S2), while the contacts Lig(ACC)–Rec(DON) retain their favorability, unlike what we saw before (see Figure S1). At the same time, the interaction energy of polar–HYD contacts, which also includes uncompensated desolvation of polar parts, is very favorable.

The reason is the highly linearly correlated CSA values between different protein atom types and a single ligand atom type (see Table S3). The correlation itself is due to the peculiarities of the binding site pockets, which are heterogeneous and contain closely spaced atoms of different types. This means that the ligand inevitably forms both complementary and non-complementary contacts with the protein.

All of the above leads us to the conclusion that non-complementary contacts at this level of detail do not fulfill the expected role of a counterweight to complementary contacts and their explicit consideration does not take SF predictions to a

Table 1 Statistical metrics (in scoring power test) of different linear regression models aimed at reproducing binding free energy.

Ligand–receptor interactions	R^2		MAE/kcal mol^{-1}		RMSD/kcal mol^{-1}	
	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

qualitatively new level. Moreover, due to the high degree of linear correlation between complementary and non-complementary CSAs (see Table S3), they are already implicitly taken into account by most existing SFs. Another point is that the current model does not include any terms that are shown to be unfavorable in terms of contribution to binding energy. However, the process of binding and formation of ligand–protein contacts is always accompanied, *e.g.*, by the loss of ligand–water and protein–water contacts (a penalty for desolvation), which is not taken into account by the current model and is rarely taken into account by empirical SFs.

Returning to what an experienced medicinal chemist might see beyond the predictions of current SFs, several conclusions can be drawn. First, the dominant and highly robust factor that is crucial for ligand–receptor interactions is the combination of shape complementarity[22] and HYD interactions,[15] which are both accounted for by CSA[HYD+HYD]. This fits well with both existing theory and practice of medicinal chemistry. These types of interactions have been consistently used in almost all known SFs,[15,17,23] which are in good agreement with the occurrence of close contacts in experimental structures.[16]

Second, the role of contact complementarity in terms of contact types is, somewhat surprisingly, not as important as might be expected, at least in the context of free energy predictions for relatively well-formed complex geometries taken from experimental structures. Even complementary hydrogen bonding interactions do not seem important enough to be included explicitly in order to achieve a better prediction of binding free energies. Part of the reason for this is that polar groups (both hydrogen acceptors and hydrogen donors) must first be desolvated before they can interact with each other. Evidently, aqueous media provide an equal or greater number of hydrogen bonds with each partner, which makes the overall energy gain unfeasible. The role of complementary hydrogen bonding (and more general electrostatic interactions) is most likely to provide specificity and selectivity for the binding a ligand to a specific target. Thus, complementary contacts increase the probability that interaction with any other target (hence off-target effects) will be less pronounced, since purely HYD interactions are not specific in nature. On the one hand, polar decoration is necessary to ensure the drug-like ADMET properties.[21] Therefore, complementary polar interactions are far more important at the lead optimization stage. On the other hand, it has been shown that explicit consideration of non-complementary interactions leads to a certain improvement in model statistics. The description of finer-grained contacts and most likely other factors controlling the magnitude of ligand–receptor free energy should definitely be considered, with a better description of desolvation being the most likely suspect for further study.

Third, it appears that when a drug-like molecule interacts with a biologically relevant target, several types of contacts (both complementary and non-complementary) are simultaneously established. Their mutual presence is manifested in the strong mutual correlation of the contact surfaces of each possible contact pair revealed in our work. Thus, using the principle of Occam's razor, in the early stages of developing the SF it was reasonable to leave only complementary contacts. However, as already indicated in the research devoted to such an SF as ScorpionScore,[14] it seems that it is not just the counting of contacts that is important, but their positive cooperativity. Thus, we hypothesize that an experienced computer-aided drug discovery expert would be able to distinguish such cases of positive cooperativity from cases in which cooperativity is not feasible.

Fourth, it is interesting to note that we found that the set of CASF-2016 complexes is by no means homogeneous in terms of both ligand structures and underlying ligand–receptor interactions. In particular, we show that the Q1 and Q4 subsets of complexes in terms of LE values typically require significantly different linear models, so the lack of explicit consideration of non-complementary interactions is not a major source of residual errors when describing experimental free energies using current SFs, most of which already take into account complementary hydrogen bonding interactions.

Finally, we hypothesize that the presence of non-complementary contacts *per se* is not drastic for free energy, but rather indicates missed opportunities in optimizing ligand specificity and selectivity. Therefore, we expect that in the near future there will be a special class of SFs that target the overall probability of a ligand becoming a drug, rather than just the mere affinity of the ligand *per se*.

The results obtained showed that the *a priori* prospective direction for further development of SFs turned out to be not as promising as expected. At the same time, the well-defined ligand–receptor complexes[9] from CASF-2016 that differ in LE appear to have different binding patterns, requiring sufficiently different models to describe the interactions. The above results will help develop new SFs with higher accuracy and wider applicability to be used in drug discovery.

This work was supported by the Russian Science Foundation (grant no. RSF 22-23-00729).

Online Supplementary Materials

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.mencom.2023.10.021.

References

1 D. Rognan, *Pharmacol. Ther.*, 2017, **175**, 47.
2 G. K. Kiriiri, P. M. Njogu and A. N. Mwangi, *Future Journal of Pharmaceutical Sciences*, 2020, **6**, 27.
3 J. Li, A. Fu and L. Zhang, *Interdiscip. Sci.: Comput. Life Sci.*, 2019, **11**, 320.
4 J. Liu and R. Wang, *J. Chem. Inf. Model.*, 2015, **55**, 475.
5 L. Pinzi and G. Rastelli, *Int. J. Mol. Sci.*, 2019, **20**, 4331.
6 T. Cheng, Q. Li, Z. Zhou, Y. Wang and S. H. Bryant, *AAPS J.*, 2012, **14**, 133.
7 A. Fischer, M. Smieško, M. Sellner and M. A. Lill, *J. Med. Chem.*, 2021, **64**, 2489.
8 C. Shen, J. Ding, Z. Wang, D. Cao, X. Ding and T. Hou, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2020, **10**, e1429.
9 H. Li, K.-H. Sze, G. Lu and P. J. Ballester, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2020, **10**, e1465.
10 M. Volkov, J.-A. Turk, N. Drizard, N. Martin, B. Hoffmann, Y. Gaston-Mathé and D. Rognan, *J. Med. Chem.*, 2022, **65**, 7946.
11 M. Su, G. Feng, Z. Liu, Y. Li and R. Wang, *J. Chem. Inf. Model.*, 2020, **60**, 1122.
12 Y. Li and J. Yang, *J. Chem. Inf. Model.*, 2017, **57**, 1007.
13 A. Bender, N. Schneider, M. Segler, W. P. Walters, O. Engkvist and T. Rodrigues, *Nat. Rev. Chem.*, 2022, **6**, 428.
14 B. Kuhn, J. E. Fuchs, M. Reutlinger, M. Stahl and N. R. Taylor, *J. Chem. Inf. Model.*, 2011, **51**, 3180.
15 C. Bissantz, B. Kuhn and M. Stahl, *J. Med. Chem.*, 2010, **53**, 5061.
16 R. Ferreira de Freitas and M. Schapira, *MedChemComm*, 2017, **8**, 1970.
17 M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li and R. Wang, *J. Chem. Inf. Model.*, 2019, **59**, 895.
18 Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu and R. Wang, *Bioinformatics*, 2015, **31**, 405.
19 Z. Liu, M. Su, L. Han, J. Liu, Q. Yang, Y. Li and R. Wang, *Acc. Chem. Res.*, 2017, **50**, 302.
20 R. Tibshirani, *Statistics in Medicine*, 1997, **16**, 385.
21 T. W. Johnson, R. A. Gallego and M. P. Edwards, *J. Med. Chem.*, 2018, **61**, 6401.
22 P. C. D. Hawkins, A. G. Skillman and A. Nicholls, *J. Med. Chem.*, 2007, **50**, 74.
23 S.-Y. Huang, S. Z. Grinter and X. Zou, *Phys. Chem. Chem. Phys.*, 2010, **12**, 12899.

Received: 26th July 2023; Com. 23/7217