Author/s:
Spackman, PR;Yu, LJ;Bond, CS;Spackman, MA;Jayatilaka, D;Thomas, SP;Yu, LJ;Morton, CJ;Parker, MW;Parker, MW;Thomas, SP

Title:
Bridging Crystal Engineering and Drug Discovery by Utilizing Intermolecular Interactions and Molecular Shapes in Crystals

Date:
2019-01-01

Citation:
Spackman, P. R., Yu, L. J., Bond, C. S., Spackman, M. A., Jayatilaka, D., Thomas, S. P., Yu, L. J., Morton, C. J., Parker, M. W., Parker, M. W. & Thomas, S. P. (2019). Bridging Crystal Engineering and Drug Discovery by Utilizing Intermolecular Interactions and Molecular Shapes in Crystals. Advanced Materials, 58 (47), pp.16936-16940. https://doi.org/10.1002/ANGE.201906602.

Title: Bridging crystal engineering and drug discovery by utilizing intermolecular interactions and molecular shapes in crystals

Authors: Peter R Spackman; Li-Juan Yu; Craig J. Morton; Michael W. Parker; Charles S. Bond; Mark A. Spackman; Dylan Jayatilaka; Sajesh Pynadath Thomas, Ph.D.

To be cited as: 10.1002/ange.201906602

Link to VoR: https://doi.org/10.1002/ange.201906602

Bridging crystal engineering and drug discovery by utilizing intermolecular interactions and molecular shapes in crystals

Peter R. Spackman,[a,b]* Li-Juan Yu,[a,c] Craig J. Morton,[d] Michael W. Parker,[d,e] Charles S. Bond,[a] Mark A. Spackman,[a] Dylan Jayatilaka,[a] Sajesh P. Thomas[a,f]*

Abstract: Most structure-based drug discovery protocols utilize crystal structures of receptor proteins. Crystal engineering, on the other hand, utilizes the wealth of chemical information inherent in small-molecule crystal structures in the Cambridge Structural Database (CSD). Here we show that interaction surfaces and shapes of molecules in experimentally determined small-molecule crystal structures can serve as effective tools in drug discovery. Our description of the shape and interaction propensities of molecules in their crystal structures can be used to screen them for specific binding compatibility with protein targets – as demonstrated through a high-throughput profiling of around 138,000 small molecule structures in the CSD, and a series of drug-protein crystal structures. Electron density-based intermolecular boundary surfaces in small-molecule crystal structures and in target protein pockets are utilized to identify potential ligand molecules from the CSD – based on 3D shape and intermolecular interaction matching. Bridging small-molecule crystallography and protein crystallography results in the identification of ligand molecules from the CSD with remarkable chemical diversity, promising receptor-binding indicators and suitable drug-likeness parameters.

Identifying suitable ligand molecules that bind to a target protein site is an important step in rational drug discovery. Drug-protein supramolecular chemistry may still be conceived in the simple terms of the lock and key model proposed by Fischer more than a century ago,[1] and thus ligand molecules which most effectively interact with the receptor – that best fit the receptor pockets – may be ranked as the most suitable drug candidates through structure-based virtual screening (SBVS) procedures such as docking or ligand-based virtual screening (LBVS).[2]

As the role of molecular shape (and shape complementarity with respect to the pockets in binding sites) is crucial in determining an effective inhibitor, many computer-aided drug design/discovery methods focus on profiling molecular shapes[3] and/or electrostatics.[4] For identification of 'pharmacophore' points for drug discovery, different mathematical descriptions or fingerprints associated with molecular shape have also been utilized.[5] On the other hand, the approach of crystal engineering[6] utilizes the information on intermolecular interactions in small-molecule crystals for the design of new crystals.

[a] Dr. Sajesh P Thomas, Dr. Peter R. Spackman
[a]School of Molecular Sciences, University of Western Australia, Perth, WA 6009, Australia; [b]School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom; [f]Department of Chemistry and iNano, Aarhus University, Langelandsgade 140, Aarhus 8000, Denmark.
E-mail: sajeshpthomas@gmail.com, peterspackman@fastmail.com
[b] Dr. Li-Juan Yu, Dr. Craig J. Morton, Prof. Michael W. Parker, Prof. Charles S. Bond, Prof. Mark A. Spackman, Prof. Dylan Jayatilaka
[c]Research School of Chemistry, Australian National University, Canberra, Australia. [d]Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville VIC 3010, Australia; [e]St Vincent's Institute of Medical Research, Fitz-roy VIC 3065, Australia.

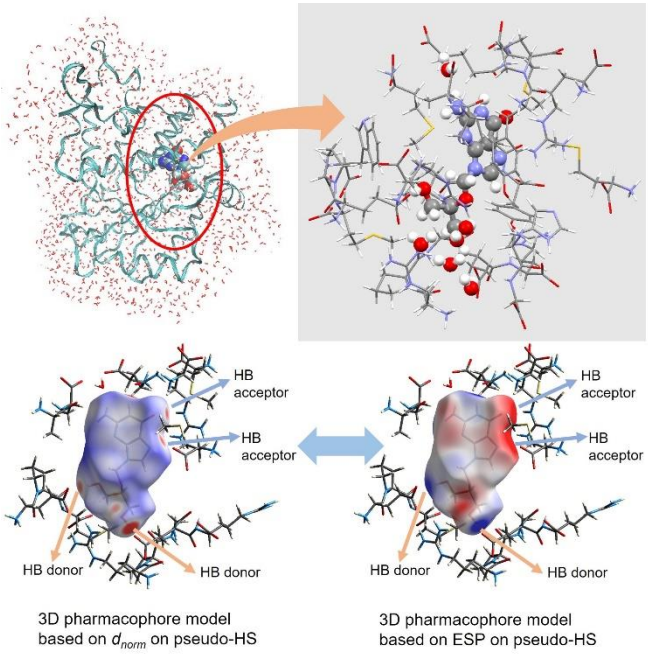


Figure 1. Top: the model ligand Ganciclovir (PDB reference code 1KI2) in its binding pocket (left) with simulated aqueous environment, and the protein fragments that directly interact with the ligands. Bottom: Proposed shape-interaction pharmacophore models based on the short contact descriptor d_{norm}-on-Hirshfeld surface (left) and the ESP-on-Hirshfeld surface (right) of the model ligand shown with some of the residues forming hydrogen bonds (HBs). While ESP surface distinguishes HB donor and HB acceptor groups, d_{norm} represents both as short contacts. Atom labels: C- grey, N-blue, O-red, S- yellow, H-white/brown.

It is known that the molecular conformations of ligands in their small-molecule crystal structures correlate well with their protein-bound conformations.[7] However, the extent to which the intermolecular interactions of these ligands in their small-molecule crystal environment mimic those in protein pockets remains largely unexplored. Earlier reports have indicated the possibility of utilizing these interactions in crystals for ligand screening[8] and for modelling target binding.[9] Effectively utilizing the wealth of chemical information in the million or more known small-molecule and protein crystals is a logical step in improving the success of early stages in the drug discovery pipeline. In this context, we report an approach that combines 3D molecular shapes and intermolecular interactions (in terms of surface properties) of small-molecule crystal structures (**Figure 1**). Our study has demonstrated a promising ligand-based screening method utilizing molecules in the Cambridge Structural Database (CSD)[10] - for the identification of suitable molecules from the CSD as potential ligands to bind specific receptor sites via an efficient search-and-match algorithm.

The CSD contains over 1 million small-molecule crystal structure entries. Several computational data mining approaches have made effective use of the vast number of experimental

small-molecule crystal structures present in the CSD, providing useful inputs towards drug design and discovery.[7a, 11] However, none of these approaches allow direct screening of the molecules in the CSD for specific receptor-binding affinities based on their molecular shape complementarity and intermolecular interactions. The combined surface-shape-property technique we present here is built upon an efficient computational method[12] that enables high-throughput profiling of molecular shapes in crystals using spherical harmonic shape descriptors of molecular surfaces, in this case the Hirshfeld surface (HS). [13] The HS not only describes the shape of a molecule, but its shape subject to the intermolecular interactions with its (typically crystalline) environment. Coupling this shape with surface properties such as electrostatic potential (ESP) or a short-contact descriptor called d_{norm} mapped on HS[13c] seems a natural extension to enhance the description of the molecular environment.

As a first step, a proof of concept test was conducted to verify the efficacy of our shape-interaction descriptor in identifying similar molecules and environments from the CSD, using eight common polymorphic drugs whose crystal structures are available in the CSD. For each of these crystal structures, the blind identification of its existing polymorph from the CSD served as a 'litmus test' for the efficacy of our description. With no explicit chemical information, purely utilizing shape-property descriptors, this screening identified polymorphs within the top-20 matches for all the eight drugs examined (CSD codes and details of the polymorph blind test is given in Supporting Information Table S2; section S2).

In order to assess the utility of such a descriptor to screen ideal ligand molecules from the CSD that can bind to specific receptor pockets, we carried out the screening in three major steps: 1) generate 3D reference models by combining the shape and interaction surfaces of the known ligand molecules inside their protein receptor pockets (as shown in **Figure 1**); 2) analyze the molecular shapes and interaction surfaces for a large subset of structures in the CSD, and store these descriptors in a library; and 3) identify potential ligand molecules from this shape-surface library that match best with the 3D pharmacophore model surface in terms of shape and surface properties - d_{norm} and ESP. While the main results are described here, the technical and computational details of our approach are given in the Supporting Information. In this study, we explored four ligand-protein structures from the Protein Data Bank (PDB)[14] to demonstrate the validity of our procedure. These ligand-protein structures are given in Table 1 along with the drug/ligand molecules (**Figure 2,** referred to as 'model ligands' hereafter) in their receptor pockets. For the generation of the reference models, we extended the idea of Hirshfeld surfaces into protein crystal structures, where the HS may be generated about the ligand molecules inside the receptor pockets (**Figure 1**). These surfaces capture the shape and interaction environment of the model-ligand in its receptor pocket, in a crude fashion. One technical challenge in generating such pseudo-HSs is that many protein sites have open pockets,making it difficult to have a closed intermolecular boundary surface covering these openings.

Table 1. Protein-ligand co-crystals examined in this study, with their reported biological activity.

Ligand #	PDB reference code	Reported activity
1	1KI2 [15]	DNA polymerase inhibitor
2	1UWH[16]	Protein kinase inhibitor
3	1Z95[17]	Nonsteroidal antiandrogen
4	3A2O[18]	HIV protease inhibitor

Since the protein pockets and the binding ligands are generally surrounded by water molecules in biological environment, we

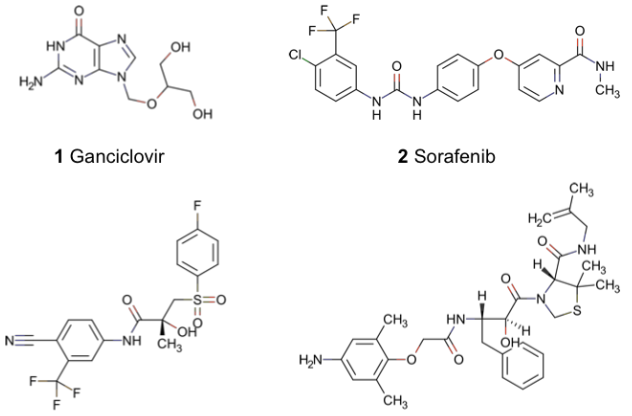


Figure 2. 2D chemical structures of the four model-ligands presented in this study.

have utilized this to generate our shape-interaction surface model. Molecular dynamics (MD) simulations were employed, with a large number of water molecules solvating the protein-ligand systems to account for the aqueous environment surrounding the proteins, and to ensure closed surfaces around the ligand molecules (See Supporting Information Section S1, Figures S1-S3). Stable protein-ligand structures in water clusters obtained from the MD simulations were then used to generate the reference models, which are in turn used as queries to be matched against the library of descriptors previously mentioned in step 2. This library consisted of descriptors of molecular shapes and intermolecular interaction (d_{norm}) for around 138,000 molecules from over 116,000 crystal structures in the CSD. The information on shape and property is stored in this library based on spherical harmonic transforms on a complex-valued function, where the real component represents the radial distribution shape of the surface and the imaginary component represents associated surface properties like d_{norm} or ESP (See Supporting Information Section S1 for the details). Subsequently, we carried out the ligand screening i.e. the matching of the surfaces and the surface property d_{norm} of reference models to their corresponding structures in our library. The best-matching molecules were identified through nearest neighbor search in the 24-dimensional coordinate space of their spherical harmonic transforms (SHT) based descriptors. For comparison (exploring the effect of environment on surface shape in our procedure) we performed an identical screening using promolecular electron density surfaces (PS) in place of HS.

To assess the effectiveness of the screening method, the top-matching molecules identified from the CSD needed to be tested for their receptor-binding affinity in comparison to the original ligands (model ligands). Therefore, we set out to examine the docking scores of the ligand candidates compared

to the reference ligands. Docking calculations performed on the top 100 matches for each compound with both surface types (HS and PS).

For three of the four reference models, we found candidates from the CSD which exhibited better docking scores than the model-ligands. For a representative example of the results – with a model of Ganciclovir (shown with ESP mapped on the surface) and the HS of the top-matching candidates from the CSD – see **Figure S4**. **Figure 3** shows a violin plot with the distributions of the docking scores for the top 100 matching ligands from the CSD for the screening procedures utilizing HS and PS, for the four model systems in this study. Note that in the case of sorafenib (model ligand **2**), nearly a quarter of the matching compounds showed improved docking scores compared to the query ligand (see Tables S11-S18).

Another important outcome of our screening is the matching of the reference models with known drugs in the CSD. The CSD contains roughly 2000 structures which are indicated to be drugs, of which around 500 were contained in the subset library explored here. Notably, 12 known drug molecules were found in the top matches during our screening for the four receptor sites (Table S1 in the Supporting Information). Their docking scores indicate that many of these drugs may exhibit similar receptor-binding to that of the model-ligands. Indeed, in the case of receptor-model 2, the top-matched ligand regorafenib (CSD code: QABCEE) is a known kinase inhibitor, and similarly for receptor-model 4, both the top-matched ligands Adefovir dipivoxil (CSD code: WUWCUO) and Indinavir (CSD code: DIJWOJ) are drugs with known antiviral activity (specifically targeting HIV). Identifying known drug molecules for novel biological targets has significant prospects in drug discovery as these compounds have already been assessed for their biocompatibility, and often may have undergone clinical trials demonstrating their safety.

Additionally, we estimated the free energy of binding for the top matching ligands in their docked positions in protein pockets, using a free energy model that has been shown to successfully predict experimental binding affinities making use of atom···atom contacts instead of residue contacts.[19] The results showed that many of the top candidates have improved binding free energies compared to the model ligands 1, 2 and 4 (see Figure 3 and Tables S3-S10). Finally, in order to assess the drug-likeness of the screened ligands in general, we examined the 800 molecules obtained as top-matches in our screening for desirable absorption or permeation based on Lipinski's empirical rules[20]. The distributions of number of H-bond donors and acceptors, molecular weights, Log P (cLogP, the logarithm of partition coefficient between n-octanol and water)[21] and polar surface area (PSA)[22] evaluated for the top-matching ligands from the CSD satisfied Lipinski's rules for drug-likeness - further substantiating the validity of the CSD based ligand screening (see Supporting Information Figures S5,S6).

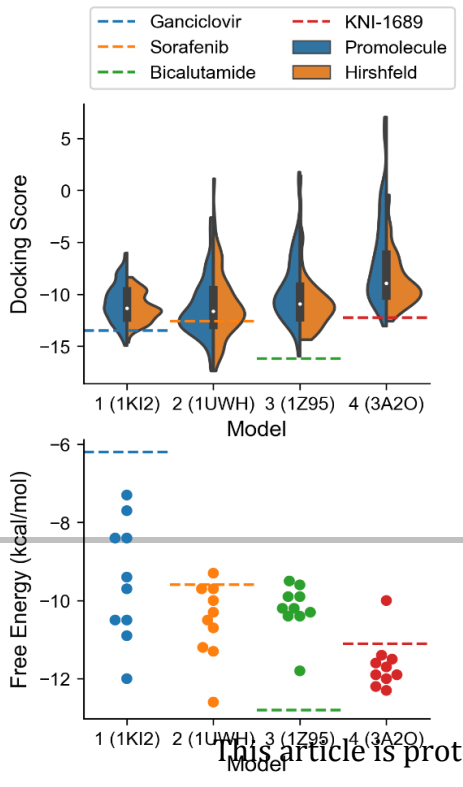


Figure 3. **Top:** Violin plot of the distributions of the docking scores for the top 100 shape-interaction (d_{norm}) matching ligands identified from the CSD, for the four model ligand-receptor systems. In the centre of each plot is a box-and-whisker plot of docking scores, and the width of each side is proportional to the number of structures with that score for the promolecule surface (left side) and the Hirshfeld surface (right). **Bottom:** Distributions of the predicted binding free energies for top 10 ligands screened from the CSD in comparison with the model ligands for each target. For reference, the docking score and free energy of the model ligands is overlaid on each plot with a dotted line.

The molecular structures of the matching ligands found in this study showed significant chemical variety, which facilitates the discovery of entirely different classes of ligands for a given receptor - as opposed to the retrospective path to design or discover new analogues by functional group/chemical modification on a current hit or lead molecule.[23] Finding several known drugs among the top-matching ligands in our screening indicates that the method may have the potential to find newer

protein-targets for known drugs. The speed of this technique lends it to application as a pre-screening method prior to docking or other established virtual screening protocols; for reference, it took mere hours to construct the library for this study, and each query for the top 100 matches took less than 1 millisecond on a personal computer. While the assessment of receptor binding of ligands based on docking scores and free energy values gives preliminary indications of binding affinities, experimental studies would of course be needed to validate the practical value of the screening outcomes. Generating target specific sub-libraries of ligand molecules using our shape-interaction screening and utilizing small molecule crystal structure libraries may help enhance the efficacy of existing drug discovery protocols. As a future extension of this work, it is possible to design idealized ligand molecules inside the protein pockets using the crystal engineering principle of 'supramolecular synthons'[24] – i.e. by adding synthon based functional groups in the model ligands for effective binding and utilize these models for screening from given libraries.

Experimental Section

Crystal structures discussed are all retrieved from the CSD and the PDB. Details of computational approaches and screening method and results are given in the Supporting Information.

Acknowledgements

This work was supported by the Australian Research Council (DP130103304) and the Danish National Research Foundation (Centre for Materials Crystallography, DNRF93). SPT acknowledges EU funding for Marie Skłodowska-Curie Individual fellowship (grant number 798633).

Author contributions: SPT and PRS conceptualized the project. SPT identified model systems and directed the studies with inputs from PRS. The manuscript was written by SPT and PRS with contributions from all authors. PRS developed and implemented all software for the combined shape-surface search method with inputs from DJ. PRS performed the calculations and analysed the results with contributions from SPT. LY performed the MD simulations and CJM performed the docking studies and the free energy estimations.

Keywords: crystal engineering • molecular recognition • noncovalent interactions • drug discovery

[1] E. Fischer, *Ber. Dtsch. Chem. Ges* **1894**, *27*, 2985-2993.
[2] A. Lavecchia, C. Di Giovanni, *Curr. Med. Chem.* **2013**, *20*, 2839-2860.
[3] a) A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain, B. Kelley, *J. Med. Chem.* **2010**, *53*, 3862-3886; b) G. M. Sastry, S. L. Dixon, W. Sherman, *J. Chem. Inf. Model.* **2011**, *51*, 2455-2466; c) J. A. Grant, M. A. Gallardo, B. T. Pickup, *J. Comput. Chem.* **1996**, *17*, 1653-1666; d) M. Hahn, *J. Chem. Inform. Comput. Sci.* **1997**, *37*, 80-86; e) P. W. Finn, G. M. Morris, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 226-241.
[4] a) E. E. Hodgkin, W. G. Richards, *Int. J. Quantum Chem* **1987**, *32*, 105-110; b) G. Buchbauer, A. Klinsky, P. Weiß-Greiler, P. Wolschann, *J. Mol. Model.* **2000**, *6*, 425-432; c) M. J. Vainio, J. S. Puranen, M. S. Johnson, *J. Chem. Inf. Model.* **2009**, *49*, 492-502.
[5] J. O. Ebalunode, Z. Ouyang, J. Liang, W. Zheng, *J. Chem. Inf. Model.* **2008**, *48*, 889-901.
[6] G. R. Desiraju, *J. Am. Chem. Soc.* **2013**, *135*, 9952-9967.
[7] a) C. R. Groom, J. C. Cole, *Acta Cryst. D* **2017**, *73*, 240; b) H.-J. Böhm, G. Klebe, *Angew. Chem. Int. Ed.* **1996**, *35*, 2588; c) A. J. Cruz-Cabeza, J. W. Liebeschuetz, F. H. Allen, *CrystEngComm* **2012**, *14*, 6797.
[8] a) V. Aparna, G. Rambabu, S. K. Panigrahi, J. A. R. P. Sarma, G. R. Desiraju, *J. Chem. Inf. Model.* **2005**, *45*, 725-738; b) G. R. Desiraju, *Chem. Commun.* **2005**, 2995-3001; c) S. Sarkhel, G. R. Desiraju, *Proteins: Structure, Function, and Bioinformatics* **2004**, *54*, 247-259.
[9] M. W. Shi, A. N. Sobolev, T. Schirmeister, B. Engels, T. C. Schmidt, P. Luger, S. Mebs, B. Dittrich, Y.-S. Chen, J. M. Bąk, D. Jayatilaka, C. S. Bond, M. J. Turner, S. G. Stewart, M. A. Spackman, S. Grabowsky, *New J. Chem.* **2015**, *39*, 1628-1633.
[10] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Cryst. B* **2016**, *72*, 171-179.
[11] a) I. J. Bruno, J. C. Cole, J. P. M. Lommerse, R. S. Rowland, R. Taylor, M. L. Verdonk, *J. Comput. Aided Mol. Des.* **1997**, *11*, 525; b) C. Bissantz, B. Kuhn, M. Stahl, *J. Med. Chem.* **2010**, *53*, 5061; c) I. J. Bruno, J. C. Cole, M. Kessler, J. Luo, W. D. S. Motherwell, L. H. Purkis, B. R. Smith, R. Taylor, R. I. Cooper, S. E. Harris, A. G. Orpen, *J. Chem. Inform. Comput. Sci.* **2004**, *44*, 2133; d) G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727; e) M. L. Verdonk, J. C. Cole, P. Watson, V. Gillet, P. Willett, *J. Mol. Biol.* **2001**, *307*, 841.
[12] P. R. Spackman, S. P. Thomas, D. Jayatilaka, *Sci. Rep.* **2016**, *6*, 22204.
[13] a) M. A. Spackman, P. G. Byrom, *Chem. Phys. Lett.* **1997**, *267*, 215; b) J. J. McKinnon, M. A. Spackman, A. S. Mitchell, *Acta Cryst. B* **2004**, *60*, 627; c) M. A. Spackman, J. J. McKinnon, *CrystEngComm* **2002**, *4*, 378.
[14] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235.
[15] J. N. Champness, M. S. Bennett, F. Wien, R. Visse, W. C. Summers, P. Herdewijn, E. de Clercq, T. Ostrowski, R. L. Jarvest, M. R. Sanderson, *Proteins: Struct., Funct., Bioinf.* **1998**, *32*, 350-361.
[16] P. T. Wan, M. J. Garnett, S. M. Roe, S. Lee, D. Niculescu-Duvaz, V. M. Good, C. G. Project, C. M. Jones, C. J. Marshall, C. J. Springer, *Cell* **2004**, *116*, 855-867.
[17] C. E. Bohl, W. Gao, D. D. Miller, C. E. Bell, J. T. Dalton, *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6201-6206.
[18] K. Hidaka, T. Kimura, H. M. Abdel-Rahman, J.-T. Nguyen, K. F. McDaniel, W. E. Kohlbrenner, A. Molla, M. Adachi, T. Tamada, R. Kuroki, N. Katsuki, Y. Tanaka, H. Matsumoto, J. Wang, Y. Hayashi, D. J. Kempf, Y. Kiso, *J. Med. Chem.* **2009**, *52*, 7604-7617.
[19] A. Vangone, J. Schaarschmidt, P. Koukos, C. Geng, N. Citro, M. E. Trellet, L. C. Xue, A. M. J. J. Bonvin, *Bioinformatics* **2018**, *35*, 1585.
[20] a) C. A. Lipinski, *Drug Discov. Today Technol.* **2004**, *1*, 337; b) C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, **.Adv***Drug Deliv. Rev.* **2001**, *46*, 3.
[21] R. Mannhold, G. I. Poda, C. Ostermann, I. V. Tetko, *J. Pharm. Sci.* **2009**, *98*, 861-893.
[22] P. Ertl, B. Rohde, P. Selzer, *J. Med. Chem.* **2000**, *43*, 3714-3717.
[23] A. T. Johnson, *J. Chem. Educ.* **2015**, *92*, 836-842.
[24] G. R. Desiraju, *Angew Chem Int Edit* **1995**, *34*, 2311-2327.

Entry for the Table of Contents (Please choose one layout)

Layout 1:

COMMUNICATION

((Insert TOC Graphic here))

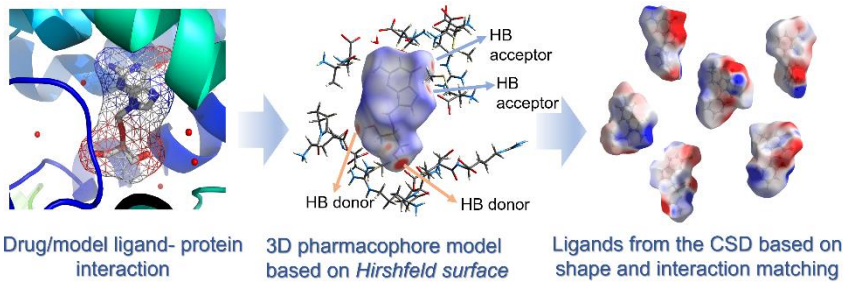
*Author(s), Corresponding Author(s)**

Page No. – Page No.

Title

Layout 2:

COMMUNICATION



Peter R. Spackman Li-Juan Yu, Craig J. Morton, Michael W. Parker, Charles S. Bond, Mark A. Spackman, Dylan Jayatilaka, Sajesh P. Thomas**

Page No. – Page No.

Bridging crystal engineering and drug discovery by utilizing intermolecular interactions and molecular shapes in crystals

Highly efficient profiling of intermolecular interactions and molecular shape in crystal structures has led to a new drug discovery approach that bridges small-molecule crystallography and protein crystallography.