

New insights and innovation from a million crystal structures in the Cambridge Structural Database

Cite as: *Struct. Dyn.* **6**, 054301 (2019); doi: [10.1063/1.5116878](https://doi.org/10.1063/1.5116878)

Submitted: 28 June 2019 · Accepted: 13 August 2019 ·

Published Online: 28 August 2019



View Online



Export Citation



CrossMark

Jason C. Cole,  Seth Wiggin,  and Francesca Stanzione 

AFFILIATIONS

The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

Note: This article is part of the Special Issue, Transactions from the 68th Annual Meeting of the American Crystallographic Association: Shining a Light on Structure-Based Drug Design.

ABSTRACT

The Cambridge Structural Database (CSD) is the world's largest and most comprehensive collection of organic, organometallic, and metal-organic crystal structure information. Analyses using the data have wide impact across the chemical sciences in allowing understanding of structural preferences. In this short review, we illustrate the more common methods by which CSD data influence molecular design. We show how more data could lead to more refined insights into the future using a simple example of trifluoromethylphenyl fragments, highlighting how with sufficient data one can build a reasonable model of geometric change in a chemical fragment with torsional rotation, and show some recent examples where the CSD has been used in conjunction with other methods to provide design ideas and more computationally tractable workflows for derivation of useful insights into structural design.

© 2019 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/1.5116878>

INTRODUCTION

The Cambridge Structural Database (CSD)¹ is a large collection of crystal structures; a recent milestone passed in June 2019 was the release of the one millionth structure to the community,² an N-heterocycle synthesized by chalcogen-chalcogen bonding catalysis.³ The large resource of structures has had an impact since the Cambridge Crystallographic Data Centre's (CCDC) inception in 1965, but the wealth of information now available has the potential to be more transformative in the coming years. Of the top 200 pharmaceutical products in 2018, 124 are small molecule compounds.^{4,5} Of these, 70 have an exact match to a crystal structure in the CSD. Small molecule structures are generally more precise and more accurate than protein structures due to their higher resolution, which allows users to gain detailed insights into molecular geometry and molecular interactions.

In this short review, we show some examples where the CSD is used in drug discovery, demonstrate a simple example where the additional information now available allows additional insight, discuss new methods to interrogate the CSD, and highlight some examples where access to this large resource is allowing innovation through screening and machine learning in other fields. In particular, we show an example where sufficient data mean we can now understand how PhCF₃ fragment's valence angles are likely to change as a function of torsional rotation; a process we might expect to see in the dynamic motion of such fragments.

HOW THE CSD IS APPLIED IN DRUG DISCOVERY

The most common use of the CSD in drug discovery is in the analysis of conformation. Brameld and co-workers have published an excellent and comprehensive review on using conformational information in drug discovery highlighting how CSD information can be very useful in making design decisions.⁶ In one example highlighted in this paper, the authors show how understanding conformational preferences can be used to optimize the binding of an inosine monophosphate dehydrogenase (IMPDH) inhibitor by understanding geometric strain (see Fig. 1). Mogul⁷ allows easy analysis of the torsional preferences and can be used to help make informed decisions on molecular design to reduce internal molecular strain. Such examples are frequent in the medicinal chemistry literature. For example, the CSD has been used to analyze substituent effects in benzamides,⁸ in the design of selective benzoxazepin PI3K δ inhibitors,⁹ and in the identification of a selective, nonprostanoid EP2 receptor agonist.¹⁰

The CSD can also be used to understand intermolecular interactions. The IsoStar¹¹ database contains information about a wealth of interactions in both the CSD and the Protein Data Bank¹² (PDB). Using the information in IsoStar can allow users to rationalize changes in affinity due to contacts within protein ligand systems. In one example, Certal and co-workers¹³ rationalized an increase in affinity on binding due to N \cdots S contacts in the observed protein ligand complex.

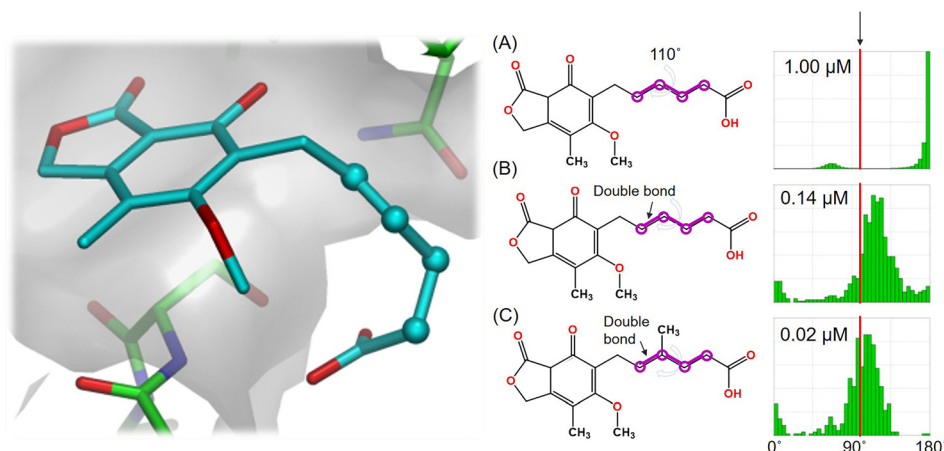


FIG. 1. Inosine monophosphate dehydrogenase (IMPDH) binder optimization. The original ligand (a) binds with a torsion angle of 110° . On the right, the observed distributions of similar torsion angles in the CSD are shown (taken from Mogul). By chemical change (going from A to B to C), one can see that the observed torsion angle (in red) is better aligned with CSD observations. The consequence is reduced strain in the inhibitor and increased bioactivity (shown as IC_{50} values of binding to inosine IMPDH).

They found that the contact was more frequent than one would expect by chance based on observations from the CSD. Intramolecular interactions too can be studied; for example, Kuhn and co-workers published a seminal review of intramolecular hydrogen bonding for medicinal chemists based (in part) on observations taken from the CSD.¹⁴

Finally, information in the CSD can be used as a data source for knowledge-based predictive algorithms. For example, the CSD has been used in various approaches for the generation of conformational ensembles, which are of general utility.^{15,16} Similarly, scoring functions for molecular docking have been derived by the analysis of interactions in the CSD.¹⁷

THE BENEFIT OF MORE INFORMATION: A SIMPLE CASE STUDY OF THE CONFORMATION OF TRIFLUOROMETHYL GROUPS BOUND TO PHENYL RINGS

Historically, users of the CSD have used informatics to interpret and understand conformational behavior. More data allow analyses that can reveal more detail. By way of example, we can take a simple case to illustrate how more data allow higher confidence and deeper insight into structural trends in crystalline systems.

Trifluoromethylphenyl groups occur more frequently in small molecule crystal structures than they have historically (the CSD v5.40 May 2019 contains 2043 structures with such a group; only 298 of these were in publications predating 2006 in the 382 652 structures

that were then available. The remaining 1745 have occurred since then in the subsequent 626 489 structures). Crystallographers are well aware that such groups are often disordered within a lattice, and indeed often occupy multiple conformations within the solid state [see, for example, the structures of both polymorphs of Leflunomide (CSD refcode family VIFQIL,¹⁸ shown in Fig. 2) both show rotational disorder around CF_3 groups within the lattice]. Higher quality structures (organic, not disordered, single crystal structures with an R-factor $<5\%$) are far rarer: only 276 structures from the 2043 structures with only 13 structures occurring before 2005 in this set.

Now we have significantly more data in the CSD, we can undertake more detailed analysis of the nature of such fragments based on higher quality structures. In the case of trifluoromethylphenyl, we can use the wealth of information to understand not only conformational preferences in the solid state, but also how the conformation of a CF_3 group is related to the preferred values of the valence angles within the fragment.

In Fig. 3, a query is shown that uses all the data in the current version of the CSD to characterize the motions of the CF_3 group with respect to the conformation around the Ph- CF_3 bond. We can analyze multiple parameters within the fragment (see Fig. 4). What becomes apparent from a CSD analysis is, first, CF_3 groups in Ph CF_3 fragments

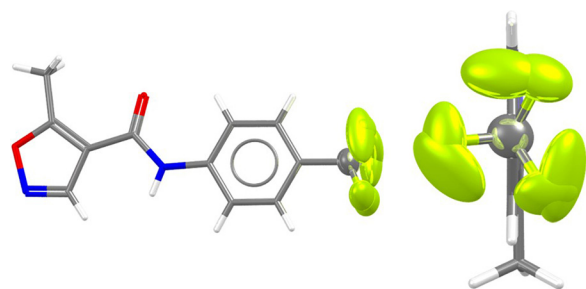


FIG. 2. Rotational disorder around the CF_3 group of polymorph II in Leflunomide (CSD refcode VIFQIL01). The CF_3 group has been refined using two alternate conformations in the lattice. The anisotropic displacement parameters, in turn, suggest additional motion around the Ph- CF_3 rotatable bond within the respective potential wells.

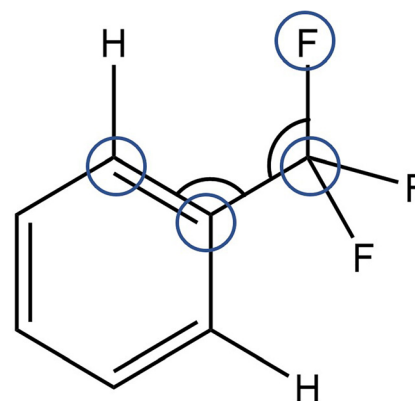


FIG. 3. Search query parameters in the CSD: F-C(sp₃)-C(ar)-C(ar) torsion angle, and F-C(sp₃)-C(ar) and C(ar)-C(ar)-C(sp₃) bond angles.

are quite conformationally free: the torsion profile shows only a slight tendency toward any given torsion angle.

Some CF_3 groups have a fluorine atom in the plane of the phenyl ring. These fluorine atoms are relatively close to a proximal hydrogen atom on the phenyl ring causing angular distortion in the plane of

the aromatic ring. The $\text{C}(\text{ar})\text{-C}(\text{ar})\text{-C}$ valence angle flexes most significantly, but in addition we see an additional distortion of the $\text{C}(\text{ar})\text{-C-F}$ angle. Both are larger to relieve the $\text{F}\cdots\text{H}$ clash.

Figure 4 also highlights the observed distributions for structures published before 2005. While the plot undoubtedly has the same

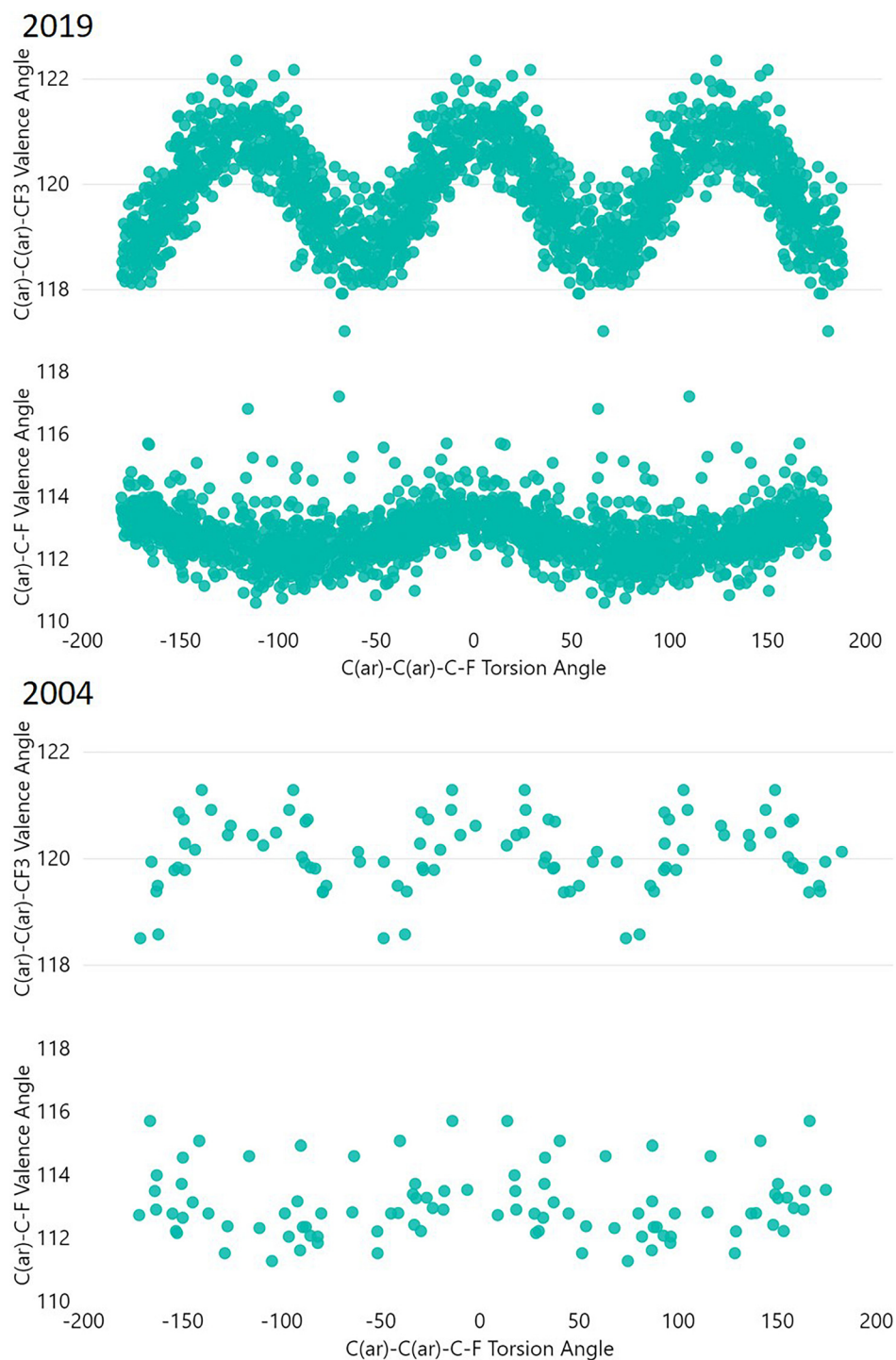


FIG. 4. Variation of internal valence angles in PhCF_3 groups with torsional rotation. At the top, results for all high precision structures are shown (organic, R-factor < 5%; no errors), at the bottom results are shown for structures up to 2004. All permutationally equivalent observations generatable from each detected fragment are included.

trends, the sparsity of data would have led to a less certain conclusion being drawn in 2004; in 2004, we could have concluded that the CF_3 torsion angle shows no strong conformational preference in the solid state. We can see suggestions toward how the internal angles vary with CF_3 rotation too in the 2004 plot, but in 2019 we can understand how the conformation occupied influences the internal angles within the PhCF_3 system in far more detail.

NEW WAYS OF SEARCHING THE CSD

CSD-CrossMiner

A wealth of data requires powerful methods for searching. CCDC has provided software systems to search and analyze information in the CSD.^{19,20} Most recently, effort has been made to provide newer methods for searching, including more elaborate pattern searching using a pharmacophore-like-representation of information within both the Protein Data Bank (PDB)^{12,21} and the CSD.²² CSD-CrossMiner is a powerful method for interactively searching based on predefined features. The method allows for searching of 3D geometric arrangements of features based SMARTS-pattern²³ feature definitions.

Figure 5 shows an example taken from a recent showcasing white paper,²⁴ showing a fairly typical query. CSD-CrossMiner allows searching of the CSD based on a more abstract representation of chemistry that is more representative of traditional medicinal chemistry thinking of pharmacophores. In addition to the built-in features, the user can define their own more features using SMARTS patterns.

Searches using CSD-CrossMiner can be used to identify structurally similar pockets in proteins. The white paper shows in addition how the software can be used for understanding cross-reactivity, finding new scaffolds based on 3D information in the CSD and to find possible bioisosteric replacements. The software has been used in pharmaceutical compound design projects;²⁵ the value of the ability to query interaction patterns for informing fragment based discovery has also been noted.²⁶

Programmatic access via a Python application programming interface

Another recent addition to the suite of methods for searching the CSD has been a Python based application programming interface (API).²⁷ The API allows versatile searching of the CSD as end users can create customized scripts. The ability to access the data via scripts in tandem with other packages such as RDKit²⁸ is very convenient for more advanced analysis of structural data as indicated by several recent examples.^{16,29–34}

Researchers have been able to develop a very useful subset of the CSD where the molecules were deemed druglike.³⁵

This drug subset in turn facilitates further analysis and comparison with the full CSD.³⁶ For example, the molecules in the subset typically have a lower formula weight than all organic molecules from the CSD. Comparison of the number of hydrogen bond donors and acceptors for an entry in the subset compared to an “average” organic CSD entry is also informative; it shows that a smaller proportion of the druglike molecules in the CSD have no hydrogen bond donors or acceptors. It also shows that fewer druglike molecules are observed with large numbers of donors or acceptors, broadly agreeing with Lipinski’s rules³⁷ in this area. Differences in the elemental composition can also be seen, with druglike molecules less likely to contain phosphorus and also favoring lighter halogens (F and Cl over Br and I). It is also interesting to observe changing trends within the druglike molecules deposited in the CSD over time. It can be seen over the past thirty years that an increasing number of structures are multicomponent (either cocrystals or salts), with the percentage of single component structures dropping from around 55% in the early 1990s to less than 40% today (see Fig. 6).

Many research projects are benefitting from API access: For example, users have been able to more easily use machine learning in tandem with the API for solvate prediction,³⁸ to help implement fragment pocket analysis using structural informatics,³⁹ and to aid with crystal structure prediction,²⁹ for understanding of the impact of compression of cocrystals⁴⁰ (of interest in the formation of tablets) and for parametrization of structural refinement programs.⁴¹

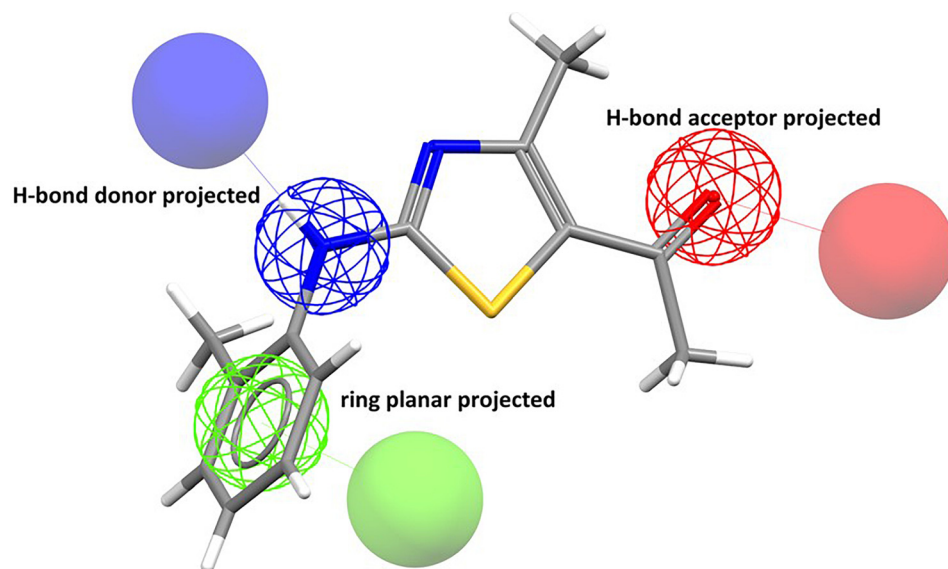


FIG. 5. A typical pharmacophore search query in CSD-CrossMiner.

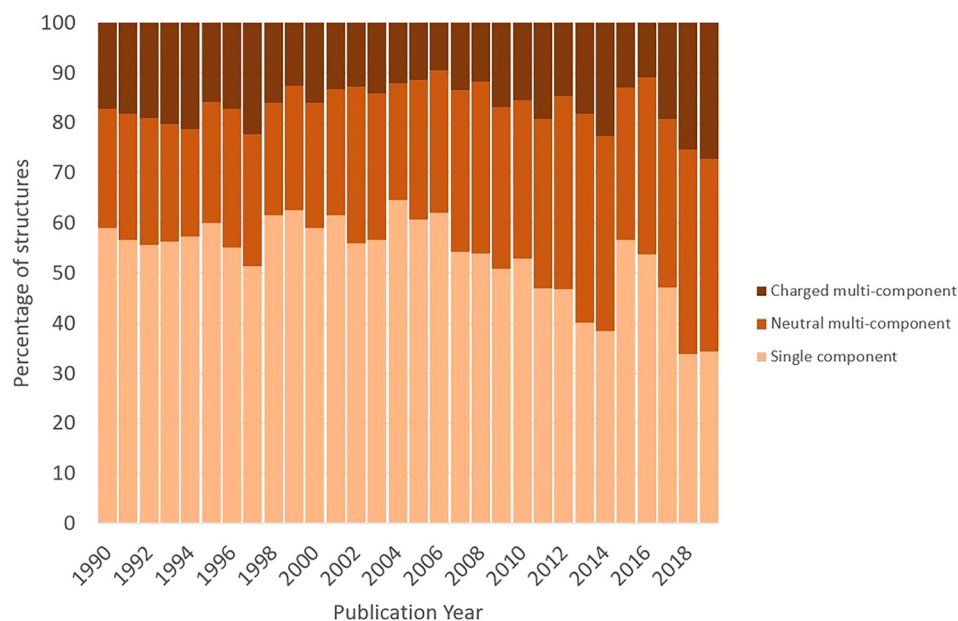


FIG. 6. How multicomponent crystalline formulation has changed with time.

Looking to the future: Opportunities for the community

The CSD contains over a million structures and continues to grow. The plethora of data means users have new opportunities available to them. We have noted that this, combined with the programmatic access to the data now available, and cheap computational power is leading to studies that would not have been tractable in the past. For example, a recent study showed how end users could effectively apply a virtual screen of the CSD to find potential high carrier mobility organic semiconductors.⁴² In this study, the authors combined data mining with various levels of quantum theory calculations to mine the CSD and find promising “pre-existing” compounds, developed for use in other areas of chemistry, that may in fact act as good candidates in this space.

The CSD can be regarded as a “big data” resource, and as such there is renewed interest in making use of the information in the CSD to solve complex problems. One interesting example of machine learning in tandem with CSD data and quantum mechanical calculations was undertaken to try to create a rapid prediction mechanism for solid state Nuclear Magnetic Resonance (NMR) shifts.⁴³ In this research, the authors used a set of 2000 diverse structures in the CSD along with solid state Quantum Mechanics (QM) based calculations (using GIPAW⁴⁴) to create a training set to train a gaussian process regression model for prediction of solid state NMR shifts. The model performs acceptably with test systems but is between 4 and 5 orders of magnitude quicker than using full QM calculations.

We should expect more machine-learned models of this type that will facilitate more rapid analysis of the solid state.

One opportunity and challenge for the community of users will be the need for more meta-data associated with structures, as such additional data will facilitate more data driven predictive modeling. Some authors are already approaching this challenge with text mining for annotation of metal-organic framework structures.³⁰

The CCDC is working toward increasing the volume of meta-data associated with structures. Two notable recent changes are the inclusion of atomic displacement parameters, which aid structural interpretation, and the inclusion of the structure factors when provided by depositors. Such information has the potential for aiding validation, but in addition may be useful for prospective analysis. In addition, depositors can now link to raw crystal structure data by including a data document object identifier during deposition. Rhetorically, we can wonder what hidden insights may be available to a researcher prepared to return to the raw crystal structure data in the future?

CONCLUSIONS

The CSD has grown to a remarkable one million structures since its inception in 1965. These structures have had a profound impact across the community, with significant impact in drug discovery and drug development. The chemical coverage of compounds in the CSD increases year-on-year as new classes of compounds are synthesized and crystallized. As the volume of data has increased more detailed insights from data have become discernible. We look forward to the next million structures and the insights they will provide.

SUPPLEMENTARY MATERIAL

See the [supplementary material](#) for underlying individual search results generated by the ConQuest search to generate the data points in Fig. 3.

REFERENCES

- ¹C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, *Acta Crystallographica Sec. B* **B72**, 171–179 (2016).
- ²See <https://www.ccdc.cam.ac.uk/News/List/the-cambridge-structural-database-reaches-one-million/> for big data leads the way for structural chemistry—The Cambridge Crystallographic Data Centre (CCDC) (last accessed June 24, 2019).

- ³W. Wang, H. Zhu, S. Liu, Z. Zhao, L. Zhang, J. Hao, and Y. Wang, "Chalcogen–chalcogen bonding catalysis enables assembly of discrete molecules," *J. Am. Chem. Soc.* **141**(23), 9175–9179 (2019).
- ⁴N. A. McGrath, M. Brichacek, and J. T. Njardarson, "A graphical journey of innovative organic architectures that have improved our lives," *J. Chem. Educ.* **87**(12), 1348–1349 (2010).
- ⁵See <https://njardarson.lab.arizona.edu/content/top-pharmaceuticals-poster> for "Top Pharmaceuticals Poster | Njardarson" (last accessed August 2, 2019).
- ⁶K. A. Brameld, B. Kuhn, D. C. Reuter, and M. Stahl, "Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis," *J. Chem. Inf. Model.* **48**(1), 1–24 (2008).
- ⁷I. 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 *et al.*, "Retrieval of crystallographically-derived molecular geometry information," *J. Chem. Inf. Comput. Sci.* **44**(6), 2133–2144 (2004).
- ⁸P.-P. Kung, E. Rui, S. Bergqvist, P. Bingham, J. Braganza, M. Collins, M. Cui, W. Diehl, D. Dinh, C. Fan *et al.*, "Design and synthesis of pyridone-containing 3,4-dihydroisoquinoline-1(2H)-ones as a novel class of enhancer of zeste homolog 2 (EZH2) inhibitors," *J. Med. Chem.* **59**(18), 8306–8325 (2016).
- ⁹B. S. Safina, R. L. Elliott, A. K. Forrest, R. A. Heald, J. M. Murray, J. Nonomiya, J. Pang, L. Salphati, E. M. Seward, S. T. Staben *et al.*, "Design of selective benzoxazepin PI3K δ inhibitors through control of dihedral angles," *ACS Med. Chem. Lett.* **8**(9), 936–940 (2017).
- ¹⁰R. Iwamura, M. Tanaka, E. Okanari, T. Kirihara, N. Odani-Kawabata, N. Shams, and K. Yoneda, "Identification of a selective, non-prostanoid EP2 receptor agonist for the treatment of glaucoma: Omidenepag and its prodrug omidenepag isopropyl," *J. Med. Chem.* **61**(15), 6869–6891 (2018).
- ¹¹I. J. Bruno, J. C. Cole, J. P. M. Lommerse, R. S. Rowland, R. Taylor, and M. L. Verdonk, "IsoStar: A library of information about nonbonded interactions," *J. Comput. Aided Mol. Des.* **11**(6), 525–537 (1997).
- ¹²H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide protein data bank," *Nat. Struct. Mol. Biol.* **10**(12), 980–980 (2003).
- ¹³V. Certal, F. Halley, A. Virone-Oddos, C. Delorme, A. Karlsson, A. Rak, F. Thompson, B. Filoche-Rommé, Y. El-Ahmad, J.-C. Carry *et al.*, "Discovery and optimization of new benzimidazole- and benzoxazole-pyrimidone selective PI3K β inhibitors for the treatment of phosphatase and TENsin homologue (PTEN)-deficient cancers," *J. Med. Chem.* **55**(10), 4788–4805 (2012).
- ¹⁴B. Kuhn, P. Mohr, and M. Stahl, "Intramolecular hydrogen bonding in medicinal chemistry," *J. Med. Chem.* **53**(6), 2601–2611 (2010).
- ¹⁵C. Schärfer, T. Schulz-Gasch, J. Hert, L. Heinzerling, B. Schulz, T. Inhester, M. Stahl, and M. Rarey, "CONFECT: Conformations from an expert collection of torsion patterns," *ChemMedChem* **8**(10), 1690–1700 (2013).
- ¹⁶J. C. Cole, O. Korb, P. McCabe, M. G. Read, and R. Taylor, "Knowledge-based conformer generation using the Cambridge structural database," *J. Chem. Inf. Model.* **58**(3), 615–629 (2018).
- ¹⁷H. F. G. Velec, H. Gohlke, and G. Klebe, "DrugScoreCSD knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction," *J. Med. Chem.* **48**(20), 6296–6303 (2005).
- ¹⁸D. Vega, A. Petragalli, D. Fernández, and J. A. Ellena, "Polymorphism on leflunomide: Stability and crystal structures," *J. Pharm. Sci.* **95**(5), 1075–1083 (2006).
- ¹⁹R. A. Sykes, P. McCabe, F. H. Allen, G. M. Battle, I. J. Bruno, and P. A. Wood, "New software for statistical analysis of Cambridge structural database data," *J. Appl. Crystallogr.* **44**(4), 882–886 (2011).
- ²⁰I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson, and R. Taylor, "New software for searching the Cambridge structural database and visualizing crystal structures," *Acta Crystallogr., B* **58**(3), 389–397 (2002).
- ²¹F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: A computer-based archival file for macromolecular structures," *J. Mol. Biol.* **112**(3), 535–542 (1977).
- ²²O. Korb, B. Kuhn, J. Hert, N. Taylor, J. Cole, C. Groom, and M. Stahl, "Interactive and versatile navigation of structural databases," *J. Med. Chem.* **59**(9), 4257–4266 (2016).
- ²³See <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> for "SMARTS—A language for describing molecular patterns" (last accessed June 24, 2019).
- ²⁴F. Stanzione, I. Giangreco, and J. C. Cole, <https://www.ccdc.cam.ac.uk/whitepapers/csd-crossminer-versatile-pharmacophore-query-tool-successful-modern-drug-discovery/> for "CSD-CrossMiner: A versatile pharmacophore query tool for successful modern drug discovery" (last accessed June 24, 2019).
- ²⁵M. Giroud, J. Ivkovic, M. Martignoni, M. Fleuti, N. Trapp, W. Haap, A. Kuglstatler, J. Benz, B. Kuhn, T. Schirmeister *et al.*, "Inhibition of the cysteine protease human cathepsin L by triazine nitriles: Amide···heteroarene π -stacking interactions and chalcogen bonding in the S3 pocket," *ChemMedChem* **12**(3), 257–270 (2017).
- ²⁶F. Giordanetto, C. Jin, L. Willmore, M. Feher, and D. E. Shaw, "Fragment hits: What do they look like and how do they bind?," *J. Med. Chem.* **62**(7), 3381–3394 (2019).
- ²⁷See <https://downloads.ccdc.cam.ac.uk/documentation/API/> for "The CSD Python API—CSD Python API 2.2.0 documentation" (last accessed June 24, 2019).
- ²⁸G. Landrum *et al.*, <http://www.rdkit.org/> for "RDKit: Open-source cheminformatics" (last accessed February 2, 2016).
- ²⁹L. Iuzzolino, A. M. Reilly, P. McCabe, and S. L. Price, "Use of crystal structure informatics for defining the conformational space needed for predicting crystal structures of pharmaceutical molecules," *J. Chem. Theory Comput.* **13**(10), 5163–5171 (2017).
- ³⁰S. Park, B. Kim, S. Choi, P. G. Boyd, B. Smit, and J. Kim, "Text mining metal–organic framework papers," *J. Chem. Inf. Model.* **58**(2), 244–251 (2018).
- ³¹M. J. Bryant, A. G. P. Maloney, and R. A. Sykes, "Predicting mechanical properties of crystalline materials through topological analysis," *CrystEngComm* **20**(19), 2698–2704 (2018).
- ³²C. R. Taylor and G. M. Day, "Evaluating the energetic driving force for cocrystal formation," *Cryst. Growth Des.* **18**(2), 892–904 (2018).
- ³³B. S. Dolinar, K. Samedov, A. G. P. Maloney, R. West, V. N. Khurstalev, and I. A. Guzei, "A chiral diamine: Practical implications of a three-stereoisomer cocrystallization," *Acta Crystallogr., Sect. C* **74**, 54 (2018).
- ³⁴D. Johnston, A. Sarjeant, and S. Wiggan, <http://scripts.iucr.org/cgi-bin/paper?S0108767318096022> for "IUCr. Temperature validation using the CSD Python API" (last accessed August 2, 2019).
- ³⁵M. J. Bryant, S. N. Black, H. Blade, R. Docherty, A. G. P. Maloney, and S. C. Taylor, "The CSD drug subset: The changing chemistry and crystallography of small molecule pharmaceuticals," *J. Pharm. Sci.* **108**(5), 1655–1662 (2019).
- ³⁶See <https://www.ccdc.cam.ac.uk/Community/blog/insights-into-drug-like-compounds-from-crystal-data/> for insights into drug-like compounds from crystal data—The Cambridge Crystallographic Data Centre (CCDC) (last accessed June 21, 2019).
- ³⁷C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Adv. Drug Delivery Rev.* **46**(1), 3–26 (2001).
- ³⁸D. Xin, N. C. Gonnella, X. He, and K. Horspool, "Solvate prediction for pharmaceutical organic molecules with machine learning," *Cryst. Growth Des.* **19**(3), 1903–1911 (2019).
- ³⁹C. J. Radoux, T. S. G. Olsson, W. R. Pitt, C. R. Groom, and T. L. Blundell, "Identifying interactions that determine fragment binding at protein hotspots," *J. Med. Chem.* **59**(9), 4314–4325 (2016).
- ⁴⁰L. E. Connor, A. D. Vassileiou, G. W. Halbert, B. F. Johnston, and I. D. H. Oswald, "Structural investigation and compression of a co-crystal of indomethacin and saccharin," *CrystEngComm* **21**, 4465–4472 (2019).
- ⁴¹M. Gilski, J. Zhao, M. Kowiel, D. Brzezinski, D. H. Turner, and M. Jaskolski, "Accurate geometrical restraints for Watson–Crick base pairs," *Acta Crystallogr., Sect. B* **75**(2), 235–245 (2019).
- ⁴²C. Schober, K. Reuter, and H. Oberhofer, "Virtual screening for high carrier mobility in organic semiconductors," *J. Phys. Chem. Lett.* **7**(19), 3973–3977 (2016).
- ⁴³F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, and L. Emsley, "Chemical shifts in molecular solids by machine learning," *Nat. Commun.* **9**(1), 4501 (2018).
- ⁴⁴J. R. Yates, C. J. Pickard, and F. Mauri, "Calculation of NMR chemical shifts for extended systems using ultrasoft pseudopotentials," *Phys. Rev. B* **76**(2), 024401 (2007).