
Article:
Gillet, V. orcid.org/0000-0002-8403-3111, Seddon, M.P. and Cosgrove, D.A. (2018) Bioisosteric replacements extracted from high quality structures in the protein databank. ChemMedChem, 13 (6). pp. 607-613. ISSN 1860-7179

https://doi.org/10.1002/cmdc.201700679

CHEMMEDCHEM

CHEMISTRY ENABLING DRUG DISCOVERY

Accepted Article

Title: Bioisosteric Replacements Extracted from High Quality Structures in the Protein Databank

Authors: Val Gillet, Matthew P. Seddon, and David A. Cosgrove

To be cited as: *ChemMedChem* 10.1002/cmdc.201700679

Link to VoR: http://dx.doi.org/10.1002/cmdc.201700679

A Journal of

ChemPubSoc Europe

WILEY-VCH

www.chemmedchem.org

10.1002/cmdc.201700679

WILEY-VCH

FULL PAPER

Bioisosteric Replacements Extracted from High Quality Structures in the Protein Databank

Matthew P. Seddon[a], David A. Cosgrove[b], Valerie J. Gillet*[a]

[a] Dr. M. P. Seddon, Prof V. J. Gillet
Information School, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK.
E-mail: v.gillet@sheffield.ac.uk
[b] Dr. D.A. Cosgrove,
Chemical Innovation Centre, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Alderley Park, UK. Current address CozChemIx Limited, Macclesfield, Cheshire, UK.

Supporting information for this article is given via a link at the end of the document.

Abstract: Bioisosterism is an important concept in the lead optimisation phase of drug discovery where the aim is to make modifications to parts of a molecule to improve some properties while maintaining others. We present an analysis of bioisosteric fragments extracted from 1458 ligands in an established data set taken from the Protein Data Bank and consisting of 121 protein targets. A pairwise analysis is carried out of all ligands for each target. The ligands are fragmented using the BRICS fragmentation scheme and a pair of fragments is deemed to be bioisosteric if the two fragments occupy a similar volume of the protein binding site. We consider two levels of generality, one which does not consider the number of attachment points in the fragments and a more restricted case in which both fragments are required to have the same number of attachments. We investigate the extent to which the bioisosteric pairs that are found are common across different targets.

Introduction

Bioisosterism is an important concept in medicinal chemistry that can be traced back to the start of the last century.[1] The lead optimisation phase of drug discovery typically involves making structural modifications, or replacements, to parts of a molecule in order to modulate some properties while maintaining others. For example, the aim might be to increase metabolic stability or solubility without affecting biological activity. Structural changes or replacements that have a neutral effect on biological function are said to be bioisosteric.[1b] Bioisosteric replacements that involve replacing the core of a compound are often referred to as scaffold hops,[2] whereas those involving replacing the substituents on a core are often referred to as functional group replacements. Traditionally, the types of bioisosteric replacements that have been explored are based on a number of well-known bioisosteres, such as those first suggested by Thornber,[3] on the experience and expertise of medicinal chemists, or on a process of trial and error. A recent review of bioisosteres in drug design is provided by Meanwell who defines bioisosteres as classical or nonclassical.[4] Classical bioisoteres are represented by structurally simple fragment pairs typified by mono-, di- and trivalent atoms or groups and ring equivalents, whereas, nonclassical bioisosteres are structurally more distinct and often comprise of different numbers of atoms and include cyclic and noncyclic isosteres. Examples of functional groups commonly found in medicinal chemistry for which there is considerable interest in identifying nonclassical bioisosteric replacements include carboxylic acid,[5] which is present in a large number of drugs but is known to have issues related to metabolic instability and toxicity, and phosphate groups which are involved in many biological processes.[6]

In recent years, significant effort has been devoted to identifying previously unknown bioisosteric replacements, to provide medicinal chemists with ideas that extend beyond their existing knowledge bases. The two main approaches have been the use of similarity methods and the application of data mining techniques to compound collections and the literature. Similarity methods are based on the similar property principle which states that similar molecules tend to have similar properties.[7] It is straight forward to extend the approach to comparing fragments and a variety of descriptors have been explored in the context of identifying bioisosteric fragments. For example, Ertl developed a web-based tool whereby a query substituent can be compared with a wide range of stored substituents based on physicochemical properties.[8] In another approach also based on physiochemical properties, Holliday et al.[9] developed the R-Group descriptor which characterises substituents by atom-based physicochemical properties based on distance from the attachment point. The properties are summed over atoms at the same topological (that is, bond) distance from the substituent connection point and represented as a vector which can then be compared with the vector derived from a query substituent. Wagener and Lommerse[10] developed a method to compare fragments based on a topological pharmacophore fingerprint that is similar in concept to atom-pairs[11] in which through bond distances between pharmacophore features are recorded. Another 2D method includes the use of reduced graph representations which also focus on pharmacophoric features.[12] A comparison of 2D fingerprint similarity and 3D molecular field similarity to screen for bioisosteres was carried out by Schuffenhauer et al.[13] who demonstrated the complementarity of the approaches.

10.1002/cmdc.201700679

Data mining methods include manual and in-silico approaches. The BIOSTER database has been compiled though manual searching of the literature for pairs of molecules that exhibit similar activities and consists of some 28,000 structural modifications.[14] The in-silico data mining approaches include matched molecular pairs (MMPs) which stem from Sheridan's work in which a maximum common substructure algorithm is used to superimpose pairs of molecules that exhibit the same activity.[15] The common substructure is removed with the remaining substituents forming the bioisosteric groups. The recently published SwissBioisostere database consists of 4.5 million replacements that have been mined from ChEMBL using MMPs.[16]

Data mining approaches have also been applied to protein-ligand complexes, which is the focus of the work described here. Watson et al. compared functional groups based on their interaction patterns to specific probe groups as observed in the Cambridge Structural Database.[17] Later, Kennewell et al. used crystallographic data to mine for target-specific bioisosteres.[18] Given a set of protein-ligand complexes for a given target, the complexes are first aligned based on the protein active sites and the aligned ligands are extracted. A reference ligand is chosen and all the ligands are fragmented. Each fragment of each ligand is then compared against the fragments in the reference and those with a high degree of volume overlap with reference fragments are assumed to have a similar role in the interaction with the target protein and defined as bioisosteric pairs. The method was used to identify target-specific bioisosteres for 12 targets taken from the Protein Databank (PDB). Recently, Zhang et al. have used the Kennewell approach to identify potential structural replacements for phosphate groups.[19] As noted earlier, these are important in medicinal chemistry for a number of reasons including the high occurrence of ATP-binding proteins, however, they are underrepresented in the SwissBioisostere database.

Other approaches based on protein ligand complexes include KRIPO and sc-PDB-Frag. In KRIPO, ligands extracted from PDB entries are fragmented with each fragment being used to define a local binding site. The local binding region is represented as a pharmacophore fingerprint and fragments that are associated with similar fingerprints, and therefore similar subpockets, are identified as potential bioisosteres. In the sc-PDB-Frag approach, bound ligands are fragmented and interaction fingerprints are calculated for each protein-fragment pair. A bioisostere is defined if a fragment pair has a low structural similarity and a high interaction pattern similarity.

Bioisosteric replacements are often considered to be generalizable; that is, it is assumed that a similar effect may be seen regardless of the particular biological activity of interest. Wassermann and Bajorath[20] recently investigated whether bioisosteric replacements could be found that are preferential for a given target. They used MMPs to identify bioisosteres from 25,000 ligands for 40 target families extracted from ChEMBL. A MMP is labelled as bioisosteric if it is found in more than one target in a family; if it occurs in different structural contexts; and if the potency difference in the ligands is moderate. Using these constraints, only 83 transformations were identified as bioisosteric, of which only 16 were found in more than one target family, with the majority, 67, being found in a single target only. As well as being target specific, bioisosteres have also been found to be context specific with respect to the structural environment in which they occur within the ligands themselves,[1a, 21] For example, in a recent review of carboxylic acid bioisosteres, Ballatore et al.[22] note the difficulty of accurately predicting the outcome of a particular isosteric replacement due to the context.

Although Wasserman and Bajorath's study is based on a large number of ligands and targets, the analysis uses 2D representations of the ligands only and does not take into account the structure of the protein. Thus, equivalences can be found between fragments which do not play similar roles when the ligands bind to their target. This is in contrast to the earlier study by Kennewell et al.[18] in which the orientation of the ligands within a binding site is taken into account and bioisosteres are defined as fragments that occupy similar volumes of the protein binding site. Kennewell's method was developed to identify target-specific bioisosteres, however, since their publication the number of structures deposited in the PDB has increased significantly thus presenting the opportunity to identify bioisosteres that are common across targets. Thus, the aim of this study is to apply Kennewell et al.'s method to a larger and more diverse set of targets for which there are protein-ligand complexes in the PDB and to consider the extent to which bioisosteric pairs can be found that are common across those targets.

Data set

The data were taken from the AstraZeneca molecular overlays for pharmacophore validation[23] which comprises of 121 overlays of high-quality crystallographic structures available to download from the Cambridge Crystallographic Data Centre. Each overlay contains a set of ligands for a particular target that have been pre-aligned based on the protein active sites. The total number of ligands available in the download is 1464, of which we were unable to process 6, leaving a total of 1458 ligands. The number of ligands per target in our set varies from 4 to 38 (with a mean number of 12). Figure 1 illustrates the overlay for the 17 ligands for the target Macrophage metalloelastase (Uniprot ID: P39900). The image shows all ligands in their three-dimensional orientation at the binding site. The full data set is shown in the Supporting Information which gives the Uniprot ID of the target protein, the target name, and the number of ligands in the target overlay.

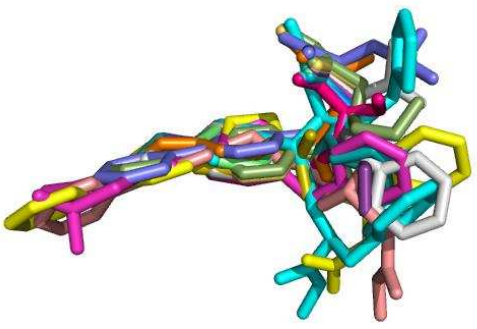


Figure 1. Overlay of ligands extracted from the target the Macrophage metalloelastase (Uniprot ID: P39900).

Methods

The methods follow those described by Kennewell et al. starting from the point at which the aligned ligands have already been extracted from the proteins, since these were directly available in the pharmacophore validation set. An overview of the algorithm used to generate bioisosteric pairs from a set of overlaid ligands for a given target site is illustrated in Figure 2. Suppose a ligand overlay contains three ligands, A, B, and C. First, ligand A is taken as the reference ligand and fragmented into a set of three fragments {a1, a2, a3}. The remaining ligands in the overlay are treated as the query ligands. For example, let B be the first query ligand, which is fragmented to produce a set of three fragments {b1, b2, b3}. The algorithm then compares the fragments of ligand A with the fragments of ligand B. All the fragments in B, are scored by their volume overlap with the first fragment in ligand A, a1. If the overlap with fragment from B, b1 for example, is greater than a given threshold then {a1,b1} forms a bioisosteric pair, as defined by Kennewell et al. As the pair of fragments is found to occupy the same volume of the active site in the receptor it is assumed that both fragments play an equivalent role and are thus labelled as bioisosteres. The algorithm then compares the remaining fragments of B against A before moving on to the fragments of C. In Figure 3 the algorithm has already moved to molecule C and found two bioisosteric pairs {a1,b1} and {a1,c2}. Once the processing of A is complete, the next molecule is taken as reference, B in this case, and its fragments compared with the fragments in all other molecules. This process may or may not result in {b1,c2} also being identified as a pair, depending on the overlap threshold that is used.

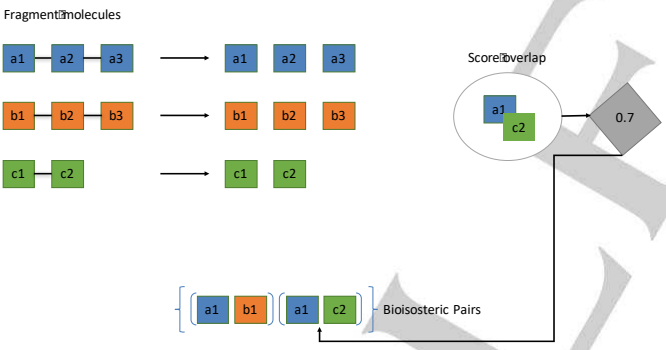


Figure 2. Illustration of the algorithm used to find bioisosteric pairs.

In our work, three different fragmentation schemes were considered initially. These are: an implementation of the fragmentation scheme from Kennewell et al. with both overlapping and non-overlapping fragments; a modification in which fragments were limited to non-overlapping fragments; and the retrosynthetic fragmentation scheme BRICS[24] with non-overlapping fragments. In Kennewell's approach, each ligand was made the reference ligand in turn and compared against all other ligands in the set. Each reference ligand was split into what the authors called "sections", which are non-overlapping fragments of the ligand. In contrast, each query ligand was fragmented into a set of overlapping fragments. Each query fragment was then compared against each section of the reference ligand. The overlapping fragments were generated by cutting single, non-ring, non-terminal bonds. In our implementation, a SMARTS representation of rotatable bonds[1] was used to search for the bonds to be broken. The bonds were then broken to form both a set of non-overlapping fragments and a set of overlapping fragments. In the case of the non-overlapping fragmentation, each molecule had all the identified bonds broken simultaneously to produce a disjoint set of fragments that constituted the original ligand, as illustrated in Figure 3(a). A set of overlapping fragments was generated by breaking one bond at a time to produce two fragments and repeating this recursively until there were no remaining bonds to break. The collection of fragments generated forms the set of overlapping fragments for each ligand, Figure 3(b).

BRICS fragmentation builds on the popular RECAP fragmentation method and uses fragmentation rules that are based on retrosynthesis,[25] with the aim of producing synthetically relevant fragments. The BRICS fragmentation scheme uses a set of 13 rules modified from the original RECAP set to obtain a more diverse fragment space.[24] The RDKit implementation of the BRICS scheme was used with the 13 rules encoded as SMARTS expressions and used to generate non-overlapping fragments. The non-overlapping fragments found using BRICS are illustrated in Figure 3(c).

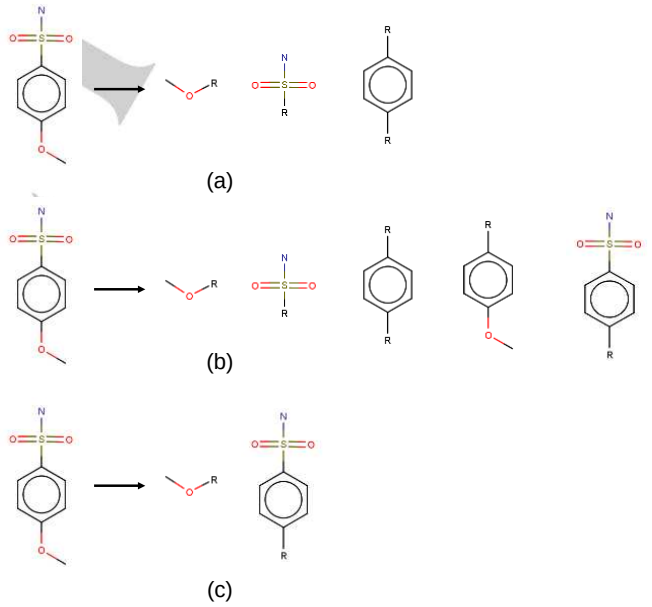


Figure 3. A ligand from target P39900 that has been fragmented according to (a) Kennewell's non-overlapping fragmentation scheme; (b) Kennewell's overlapping fragmentation scheme; and (c) the BRICS non-overlapping fragmentation scheme.

A comparison of the three fragmentation schemes shows that the overlapping method produces the largest number of fragments and may therefore give the greatest opportunity for finding bioisoteric pairs, however, it was thought that these fragments do

[1] The SMARTS used was '[!$([NH]!@C(=O))&!D1&!$(*#*)]-&!@[!$([NH]!@C(=O))&!D1&!$(*#*)]'

not accord with the notion of fragments as used in lead optimisation. The BRICS retrosynthetic scheme, on the other hand, fragments molecules on bonds that can be made through known reactions so that the resultant fragments will be more meaningful to synthetic chemists. For this reason, the non-overlapping BRICS fragmentation method was adopted for the experiments reported here.

The volume overlap for each fragment pair was calculated using a simplified Gaussian function, again following Kennewell et al. For each atom, m, in the reference fragment, the Euclidean distance to each atom, n, in the query fragment is determined and a contribution to the score is calculated. The contributions are summed over all pairs of atoms and an average score calculated as shown below. The average score was calculated to account for size bias,

$$score = \frac{2}{m+n}\sum_{j=1}^{m}\sum_{i=1}^{n} e^{-d_{i,j}^2}$$

where m and n are the numbers of atoms in the reference and query fragments, respectively, and a pair of fragments with a score greater than 0.7 was kept. The analysis was conducted at two levels: one in which the number of attachments points in each fragment was not considered and also at a more restricted level whereby the number of attachment points in each fragment should be the same. Thus, at the more general level, a pair of fragments is considered bioisosteric if they occupy a similar volume within the binding site of the target, and at the more restricted level the fragments must also have the same number of attachment points to the ligands from which they have been extracted.

Results and Discussion

Table 1 summarises the results for all 121 targets in the pharmacophore validation data set. The 1458 ligands resulted in a total of 6282 fragments when summed across all targets. 55076 fragment pairs were found by first identifying the unique pairs for each target and then summing across all targets. Of these, 3212 represent bioisosteric pairs, that is, they pass the volume overlap threshold, with this number reducing to 1851 when the condition that the number of attachment points in each fragment must be equal is applied. Figure 4 gives an illustration of how these numbers were calculated. A summary of the results for each target is shown in the Supporting Information and example bioisosteres are shown in Figure 5.

Table 1. See Figure 4 for an illustration of how these numbers were calculated. Note that Number of fragments, Number of fragment pairs and Number of bioisoteric pairs are summed over all targets.

Number of targets	Number of ligands	Number of fragments	Number of fragment pairs	Number of bioisosteric pairs	Number of cross target bioisosteric pairs	Number of bioisosteric pairs – equal number of attachments	Number of cross target bioisosteric pairs - equal number of attachments
121	1458	6282	55076	3212	146	1851	107

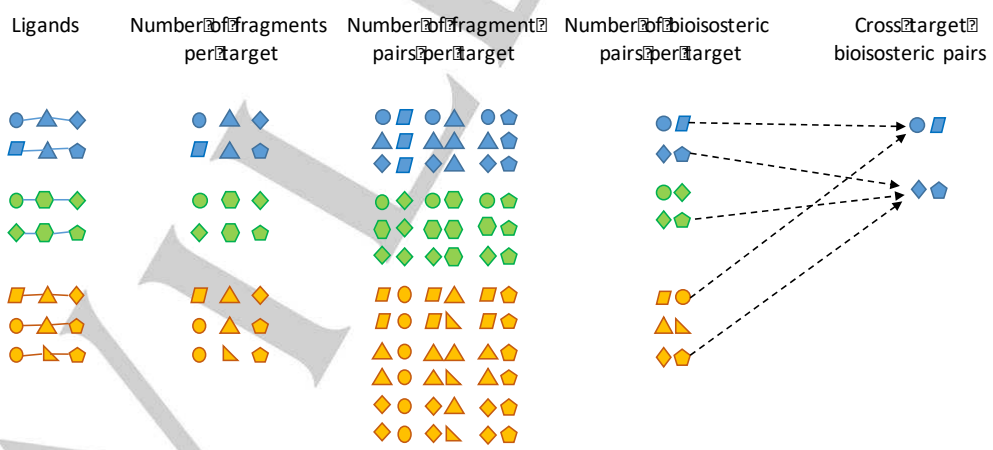


Figure 4. Schematic to illustrate how the numbers shown in Table 1 were calculated.

10.1002/cmdc.201700679

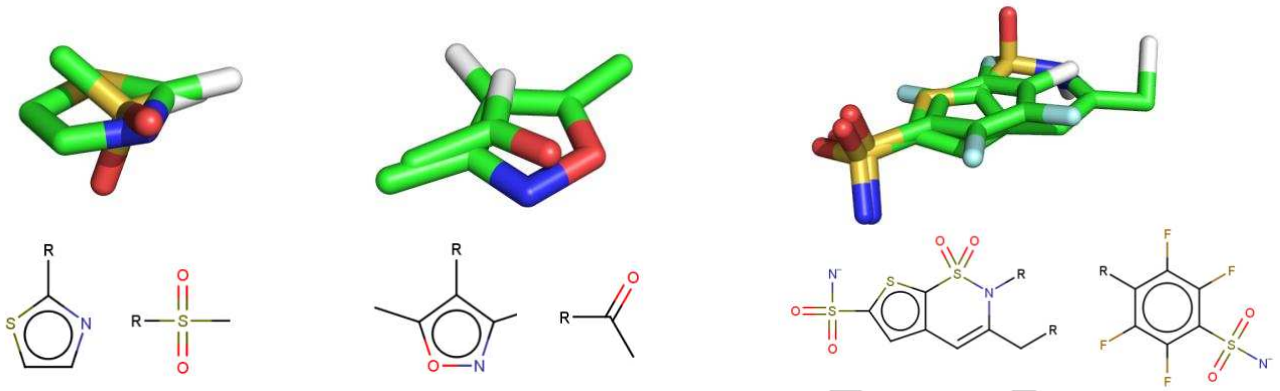


Figure 5. Example bioisosteric pairs from targets: methionine aminopeptidase (P0AE18); bromodomain-containing protein 2 (P25440); and carbonic anhydrase II (P00918).

Given the much larger number of targets in this study compared to Kennewell et al.'s, the frequency of occurrence of each bioisosteric pair across the different targets was also determined. The occurrence of a given bioisosteric pair in more than one target would provide evidence of the generalisability of the transformation beyond a given target. Bioisosteric pairs from different targets were considered the same if they had identical topological structures. A summary of the results for the more general level (that is, when the number of attachment points is not taken into account) is presented in Figure 6 where the vertical axis of the graph is the number of bioisosteric pairs that were found in a given number of targets (log scale) and the horizontal axis is the number of targets. The vast majority of the bioisosteric pairs occurred in a single target with only 146 bioisosteric pairs found in more than one target. 106 bioisosteric pairs occurred in two targets only, with this number declining rapidly: 17 were found in three targets; and only 23 were found in more than three targets. The graph can be interpreted as a measure of the generalisability of the bioisosteres found using the BRICS-fragmentation method: the horizontal axis shows the increasing generality of the bioisosteres and the vertical axis shows the extent to which the bioisosteric pairs can be generalised across a diverse number of targets. All bioisosteric pairs that were found in three or more targets are shown in Figure 7 where it can be seen that those that occur in the largest number of targets tend to be small fragments or phenyl rings with different substitution patterns. All pairs found in more than one target are shown in the Supporting Information. For the more restricted case, only 107 bioisosteric pairs were found in more than one target; 82 of these occurred in two targets; 10 were found in three targets; and only 13 were found in more than three targets.

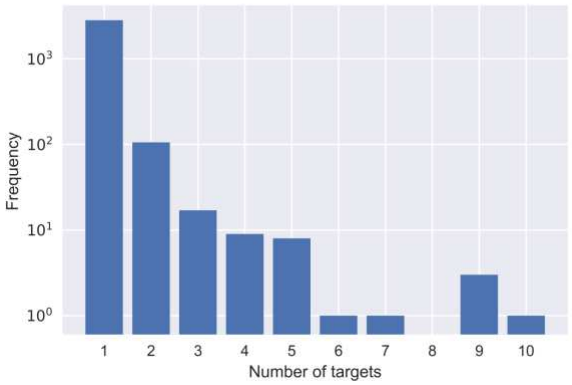


Figure 6. The frequency of occurrence of bioisosteric pairs across multiple targets.

10.1002/cmdc.201700679

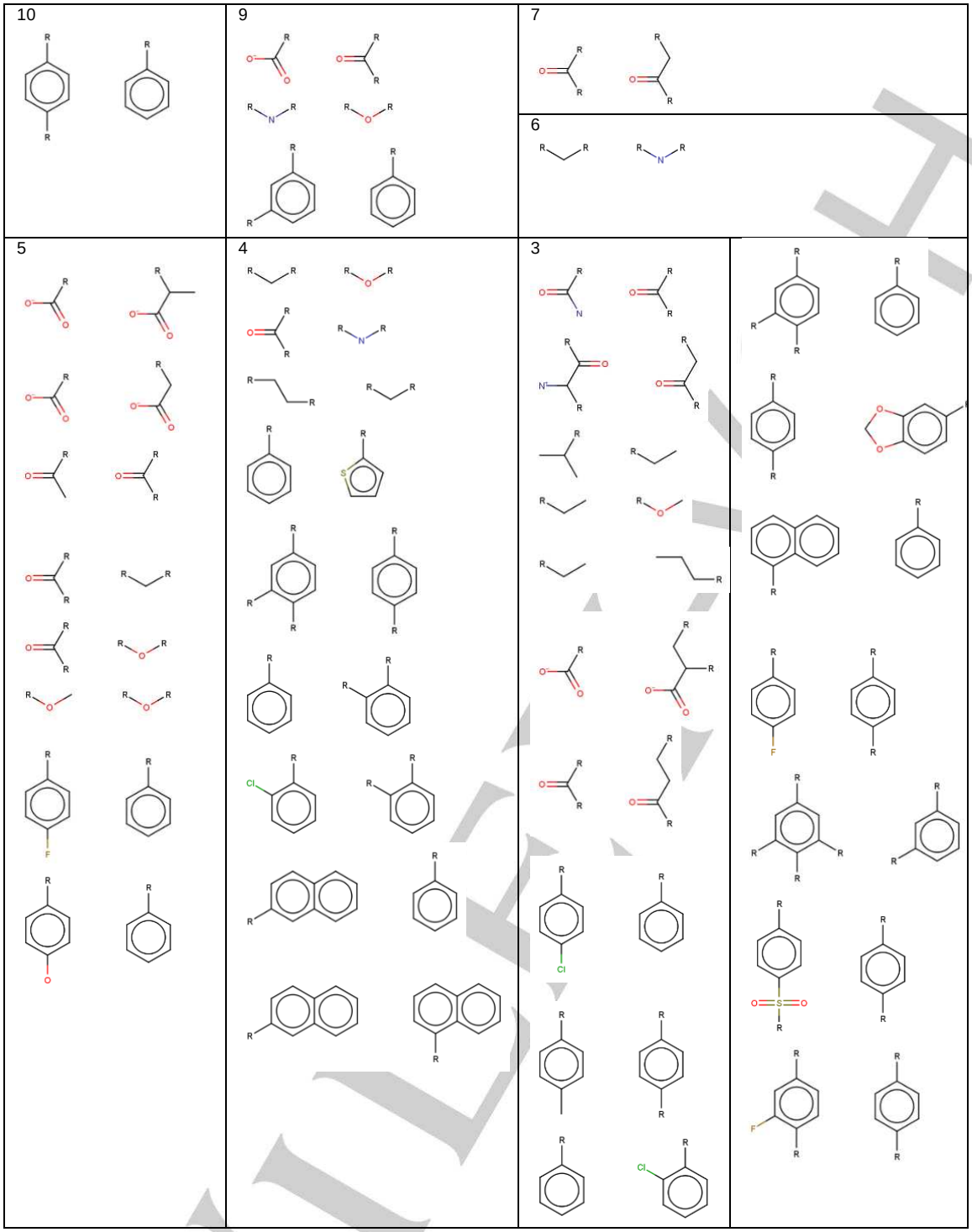


Figure 7. Bioisosteric pairs that were found in three or more targets. The number of targets is indicated in the top left corner of each cell.

Conclusions

Our results demonstrate that the vast majority of the bioisosteric pairs found using this method are unique to a single target. We found little evidence of bioisosteric pairs that are common across different targets within the PDB. Where we did find cross-target bioisosteres these were mostly common to two targets only, with those that were found in more than two targets being dominated by small classical fragments or phenyl rings with different substitution patterns. The lack of generalisability of the bioisosteres was a somewhat surprising finding given that functional group replacements are the cornerstone of lead optimisation programmes and our expectation that medicinal chemists tend to explore similar functional group replacements across different drug discovery projects. However, these results are consistent with the much larger study carried out by

Wassermann and Bajorath[20] in which the ligands and targets were extracted from ChEMBL and the analysis did not take into account the context of the target binding sites.

It should be noted that this type of analysis is highly dependent on the fragmentation scheme used. In our case we chose to use the BRICS fragmentation scheme which results in a smaller number of fragments than would be found with the approach used in the original Kennewell method. This fragmentation method can also lead to some closely related fragments (for example, that differ in the number of atoms attached to a functional group) being identified as distinct, whereas, a medicinal chemist might see these as the same functional group. These effects undoubtedly had an effect on the number of bioisosteric pairs that were found, however, we feel that this scheme is likely to produce fragment pairs that are more meaningful to synthetic chemists. One issue with the BRICS scheme is that it does not fragment fused rings which could lead to an imbalance in the sizes of fragments within a pair. Finally, and as also noted by Wassermann and Bajorath, this analysis is affected by data sparseness so that it cannot be concluded that the presence of a bioisosteric pair in only one target means that it is specific to that target. It would therefore be interesting to repeat the study in the future as new structures are added to the PDB.

Acknowledgements

We acknowledge support provided by the Biotechnology and Biological Sciences Research Council (BBSRC) and AstraZeneca via a PhD studentship awarded to MPS.

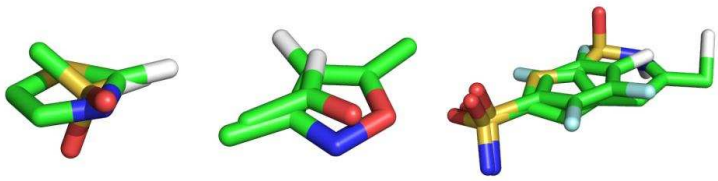
Keywords: bioisosteres • PDB • fragmentation • pharmacophore • molecular overlay

References:

[1] a) G. Papadatos, M. Alkarouri, V. J. Gillet, P. Willett, V. Kadirkamanathan, C. N. Luscombe, G. Bravi, N. J. Richmond, S. D. Pickett, J. Hussain, J. M. Pritchard, A. W. J. Cooper, S. J. F. Macdonald, J. Chem. Inf. Model. **2010**, 50, 1872-1886; b) N. Brown, Mol. Inf. **2014**, 33, 458-462.
[2] a) M. Bohl, J. Dunbar, E. M. Gifford, T. Heritage, D. J. Wild, P. Willett, D. J. Wilton, Quant. Struct.-Act. Relat. **2002**, 21, 590-597; b) S. R. Langdon, P. Ertl, N. Brown, Mol. Inf. **2010**, 29, 366-385.
[3] C. W. Thornber, Chem. Soc. Rev. **1979**, 8, 563-580.
[4] N. A. Meanwell, J. Med. Chem. **2011**, 54, 2529-2591.
[5] P. Lassalas, B. Gay, C. Lasfargeas, M. J. James, V. Tran, K. G. Vijayendran, K. R. Brunden, M. C. Kozlowski, C. J. Thomas, A. B. Smith, D. M. Huryn, C. Ballatore, J. Med. Chem. **2016**, 59, 3183-3203.
[6] C. S. Rye, J. B. Baell, Curr. Med. Chem. **2005**, 12, 3127-3141.
[7] M. A. Johnson, G. M. Maggiora, Concepts and Applications of Molecular Similarity, John Wiley & Sons, New York, **1990**.
[8] P. Ertl, J. Mol. Graphics Model. **1998**, 16, 11-13.
[9] J. D. Holliday, S. P. Jelfs, P. Willett, P. Gedeck, J. Chem. Inf. Comput. Sci. **2003**, 43, 406-411.
[10] M. Wagener, J. P. M. Lommerse, J. Chem. Inf. Model. **2006**, 46, 677-685.
[11] R. E. Carhart, D. H. Smith, R. Venkataraghavan, J. Chem. Inf. Comput. Sci. **1985**, 25, 64-73.
[12] K. Birchall, V. J. Gillet, P. Willett, P. Ducrot, C. Luttmann, J. Chem. Inf. Model. **2009**, 49, 1330-1346.
[13] A. Schuffenhauer, V. J. Gillet, P. Willett, J. Chem. Inf. Comput. Sci. **2000**, 40, 295-307.
[14] I. Ujváry, J. Hayward, in Bioisosteres in Medicinal Chemistry, Wiley-VCH Verlag GmbH & Co. KGaA, **2012**, pp. 53-74.
[15] R. P. Sheridan, J. Chem. Inf. Comput. Sci. **2002**, 42, 103-108.
[16] M. Wirth, V. Zoete, O. Michielin, W. H. B. Sauer, Nucleic Acids Res. **2013**, 41, D1137-D1143.
[17] P. Watson, P. Willett, V. J. Gillet, M. L. Verdonk, J. Comput.-Aided Mol. Des. **2001**, 15, 835-857.
[18] E. A. Kennewell, P. Willett, P. Ducrot, C. Luttmann, J. Comput.-Aided Mol. Des. **2006**, 20, 385-394.
[19] Y. Z. Zhang, A. Borrel, L. Ghemtio, L. Regad, G. B. af Gennas, A. C. Camproux, J. Yli-Kauhaluoma, H. Xhaard, J. Chem. Inf. Model. **2017**, 57, 499-516.
[20] A. M. Wassermann, J. Bajorath, Future Med. Chem. **2011**, 3, 425-436.
[21] A. M. Wassermann, J. Bajorath, MedChemComm **2011**, 2, 601-606.
[22] C. Ballatore, D. M. Huryn, A. B. Smith, ChemMedChem **2013**, 8, 385-395.
[23] I. Giangreco, D. A. Cosgrove, M. J. Packer, J. Chem. Inf. Model. **2013**, 53, 852-866.
[24] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, M. Rarey, ChemMedChem **2008**, 3, 1503-1507.
[25] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, J. Chem. Inf. Comput. Sci. **1998**, 38, 511-522.

10.1002/cmdc.201700679

Entry for the Table of Contents



A data set of high quality protein-ligand complexes spanning 121 protein targets was analysed for the presence of bioisosteric fragments. A pairwise analysis of all ligands for each target was carried out. The ligands were fragmented and a pair of fragments considered bioisosteric if they occupy a similar volume of the protein binding site. Only a small number of the bioisosteric pairs were found to be common to two or more targets.