PERSPECTIVE

Automated molecule editing in molecular design

Peter W. Kenny · Carlos A. Montanari ·
Igor M. Prokopczyk · Fernanda A. Sala ·
Geraldo Rodrigues Sartori

Received: 3 June 2013 / Accepted: 23 August 2013 / Published online: 4 September 2013

Abstract The ability to modify chemical structures in an automated and controlled manner is useful in molecular design. This Perspective introduces the MUDO molecule editor and shows how automated molecule editing can be used to standardize structures, enumerate tautomeric and ionization states, identify matched molecular pairs. Unlike its predecessor Leatherface, MUDO can also process 3D structures and this capability can be used to link non-covalently docked ligands to proteins.

Keywords Ionization · Matched molecular pair · Molecule editor · SMIRKS · Structure standardization · Tautomer

Introduction

One definition of molecular design is control of behavior of compounds and materials by manipulation of molecular properties. Molecular design, especially in drug discovery, often requires that large numbers of molecular models be evaluated. Being able to modify chemical structures in a controlled and automated manner allows views of chemistry to be imposed on large numbers of structures objectively, transparently and reproducibly [1, 2]. For example, a vendor of compounds must distinguish a carboxylic acid from its sodium salt while a medicinal chemist selecting compounds for screening needs to recognize the equivalence of the two under assay conditions. Assembly of chemical libraries from reagents, standardization of structures (e.g. nitro groups; 'de-salting') [1–3] and setting ionization and tautomer states [1, 2, 4–6] prior to virtual screening [7] all involve automated molecule editing. Relationships between structures can be revealed by applying structural transformations (e.g. convert chlorine to fluorine) and matched molecular pair analysis (MMPA) was introduced in the context of automated molecule editing [1]. Despite the broad applicability of automated molecule editing, no general-purpose tool appears to be openly available although applications offering specific functionality such as library assembly from reagent structures and structure standardization, have been in use for a number of years. This Perspective introduces MUDO (MolecUle eDitOr), which uses SMIRKS [8] to direct the molecule editing. Although MUDO is used for the examples and its source code is provided as supplementary material, the focus of this Perspective should be seen to be automated molecule editing rather than specific software.

Leatherface

The predecessor of MUDO was the Leatherface molecule editor and it is appropriate to say something about the rather haphazard development of the latter and the motivation for creating it in 1996. Leatherface was built with the Daylight toolkit [9] and the structure editing was controlled using a combination of SMARTS [10] notation and editing instructions [1]. Leatherface was originally

Electronic supplementary material The online version of this article (doi:10.1007/s10822-013-9676-0) contains supplementary material, which is available to authorized users.

P. W. Kenny (✉) · C. A. Montanari · I. M. Prokopczyk ·
F. A. Sala · G. R. Sartori
Grupo de Estudos em Química Medicinal (NEQUIMED),
Instituto de Química de São Carlos, Universidade de São Paulo,
Av. Trabalhador Sancarlense, 400, São Carlos, SP 13566-590,
Brazil
e-mail: pwk.pub.2008@gmail.com

conceived as a means with which to generate databases for scaffold-hopping [11, 12] by breaking selected (e.g. exocyclic) bonds in molecular structures and discarding uninteresting portions of disconnected structures. Substructural context is important in these applications and one would usually want to retain the carbonyl of *N*-acylpiperidine as part of the scaffold rather than break the amidic bond. Leatherface was actually little used in the role originally intended for it and the motivation for its next phase of development came from difficulties in handling tautomers that were encountered when matching pharmacophores against 3D databases. At the time (1997) pyridones were registered in the Zeneca corporate database as their hydroxypyridine tautomers. Both Unity [13] and ALADDIN [14], which had been used previously for pharmacophore-matching, allowed definition of a geometric object corresponding to the position of a hypothetical hydrogen atom bonded to the nitrogen atom in question. Unlike ALADDIN, Unity did not (at least in 1997) allow these geometric objects to be grouped with other hydrogen atoms (e.g. bonded to heteroatoms) for creation of generic hydrogen bond donor definitions. Adding the ability to modify bond order, atomic charge and number of implicit hydrogen atoms made it possible to use Leatherface to set both tautomeric and protonation states and modify mesomeric forms such as the ylid representation of the nitro group [1, 2].

Although Leatherface had evolved into a practical tool for preparation of chemical databases by this stage, it was still unable to enumerate tautomeric and protonation states in a systematic manner. Whilst it was possible to generate alternative tautomers in a limited manner, for example by protonating imidazoles, shifting the formal charge and deprotonating, this fell short of general enumeration. The motivation for introducing enumeration in 2001 was provided by the need to build databases [4, 5] for virtual screening [7] and the enhanced capability led to some unexpected benefits. Tautomers were enumerated by Leatherface as lists of canonical SMILES [15, 16] strings which were sorted prior to output so that duplicates could be removed. The first member of this sorted list could be designated as the ‘canonical tautomer’ [1, 2] which allowed easy identification of duplicate structures when rings such as pyrazole and imidazole were present. The other unexpected benefit of introducing enumeration capability was that it made it possible to identify matched molecular pairs (MMPs) in large databases of chemical structures. Although Leatherface was used as early as 1998 to identify matched molecular pairs, the presence in a molecular structure of more than a single instance of the relevant feature (e.g. chlorine linked to aromatic carbon) complicated analysis. It was only when this enumeration capability was that Leatherface came to be used for MMPA in drug discovery projects at AstraZeneca. It should be stressed that both MMPA and canonical tautomer definition were conceived after the most of the necessary functionality was already in place and neither application can be considered to have influenced the design of Leatherface. It says something about the applicability and fundamental nature of automated molecule editing when general-purpose software such as Leatherface could be used, without modification, to perform MMPA.

MUDO essentials

MUDO, which was created with the OEChem toolkit [17], edits structures by applying structural transformations defined as SMIRKS [8] and its operation is described in detail in the manual that is included in the supplementary material. MUDO has three modes of operation and the normal mode can be described as ‘search and destroy’ since each matching substructure is transformed until no more remain. For example, normal mode deprotonation of a neutral dicarboxylic acid would convert it to its dianion. In normal mode, MUDO can process structures either with or without 3D coordinates.

MUDO is also able to enumerate structural forms, although not with 3D coordinates, because enumerated forms of a structure are processed (e.g. duplicate removal) as a list of canonical SMILES [15, 16] strings. In enumeration mode, a structure is generated separately for each substructure that matches a SMIRKS transform so two monoanions would be generated from the neutral dicarboxylic acid unless they were equivalent (e.g. oxalic acid) in which case only the single monoanion would be output. This mode of operation would typically be used for MMPA where one would need to generate both *des*-chloro analogs of a structure with two chloro substituents.

The third mode of operation is the link mode which allows covalent bonds to be formed between structures. Although the motivation for creating this capability was for building models of protein ligand complexes, the general-purpose nature of the software means that it could also be used to link two docked fragments. MUDO can process structures with or without 3D coordinates in link mode.

MUDO provides an exhaustive option that causes both normal and enumeration mode editing to be performed more extensively. Using the exhaustive option in conjunction with normal mode editing repeatedly applies the full set of SMIRKS transforms until the molecular structure remains unchanged after two successive rounds of editing. This can be used to trim complex substituents in a highly controlled manner and similar functionality in Leatherface was used to implement molecular core matching (MoCoM) [18]. The exhaustive option has a different effect on enumeration mode editing in that the input structure is included in the list of structural forms when enumeration is performed exhaustively and each structure in the list is used as a starting point for enumeration until the length of the sorted list of SMILES remains unchanged after two successive rounds of editing. Exhaustive enumeration of the neutral dicarboxylic acid of the earlier example will lead to the output of the dianion and starting structure as well as the two monoanions resulting from non-exhaustive enumeration. MUDO can also be directed to output only the first structure in the sorted list as the canonical form which will usually be one of a number of tautomers.

Protonation, tautomers and structure standardization

Some of the implications of tautomerism [19–21] and protonation for cheminformatics and molecular design have already been discussed [1, 2, 22, 23]. In the context of ligands binding to proteins, exchange of protons, either with solvent or with other atoms in the same molecule, can be said to invert polarity of atoms [1, 2]. For example, ionization of a carboxylic acid replaces a neutral hydrogen bond donor with an anionic hydrogen bond acceptor. Frequently typical pK_a values and tautomeric preferences will be known for functional groups of interest and, when this is the case, a molecule editor represents a means with which to exploit this chemical knowledge for building virtual screening databases. This was the basis of an ionization and tautomer model [1] used at AstraZeneca that was applied using Leatherface and SMIRKS transforms have also been used for imposing tautomeric preferences [24]. A molecule editor also enables structure standardization, for example by converting the ylid form of a nitro group to one lacking formal charges in which nitrogen is pentavalent. Structure standardization is a particular issue when using external collections of compounds and a molecule editor allows ‘business rules’ to be applied prior to selection of compounds [1–3].

In many molecule editing applications it is sufficient to convert one substructure into another but in other situations accurate representation of compounds requires that more than one form be used. For example, pyrazoles lacking substituents on nitrogen can exist in two tautomeric forms that are equivalent in that neither can be unambiguously designated as the preferred form. Docking these pyrazoles will in general require that both tautomers be generated although in some cases one is likely to be more stable than the other [25]. The lack of an unambiguous reference structure for pyrazoles is also an issue for compound registration systems which must associate different tautomeric forms with a single entity such as a particular sample of a compound. Both MUDO and Leatherface address this problem in an identical manner by enumerating tautomeric forms as a sorted list of canonical SMILES and designating the first member of the list as the ‘canonical tautomer’. It is important to stress that term ‘canonical’ does not mean that the tautomer is more stable or somehow ‘better’ than the other tautomers. The different ways in which MUDO enumerates pyrazole tautomers are illustrated in Fig. 1.

Protonation of piperazine provides a good example with which to illustrate how a molecule editor can be used to set physiologically relevant protonation states. The pK_a values (313 K) of 9.37 and 5.02 measured for piperazine show that the predominant form of piperazine at normal physiological pH of 7.4 is the mono-cation while those measured for 1,4-dimethylpiperazine (8.06 and 3.64) suggest that mono-alkylated piperazines will protonate preferentially on the unsubstituted nitrogen [26]. This information can be used to construct a simple protonation model in which both protonated forms are generated for 1,4-dialkylpiperazines and for piperazines lacking substituents on nitrogen but only the unsubstituted nitrogen is protonated when one nitrogen atom is substituted. The use of MUDO to apply this protonation model is illustrated in Fig. 2 and the vector bindings and the associated vector bindings and SMIRKS are listed in Tables 1 and 2.

Matched molecular pairs

Leatherface provided a means with which to make specific relationships between molecular structures the focus of searches and MMPA was introduced in this context [1]. The history of associating structural modifications with changes in values of properties can be traced to the pioneering studies of Hammett [27] and medicinal chemists had been making observations like ‘a chloro substituent at C3 increases potency’ long before Leatherface had even been thought of. It was collaboration with medicinal chemists, in particular at the Zeneca Wilmington site, which suggested that Leatherface might be used to establish relationships between chemical structures. The assumption made when analyzing data in this manner is that differences in the values of a property for compounds can be predicted with greater accuracy than the values of the property for individual compounds and a matched molecular pair model can be seen as a special type of local quantitative structure activity/property relationship (QSAR/QSPR). Although MMPA is sometimes equated with Free-Wilson analysis [28], it is more accurate to describe it as the data-analytic equivalent [29] of free energy perturbation [30] using alchemical transformations [31]. Matched molecular pairs represent just one component of a chemical space paradigm in which activity and properties

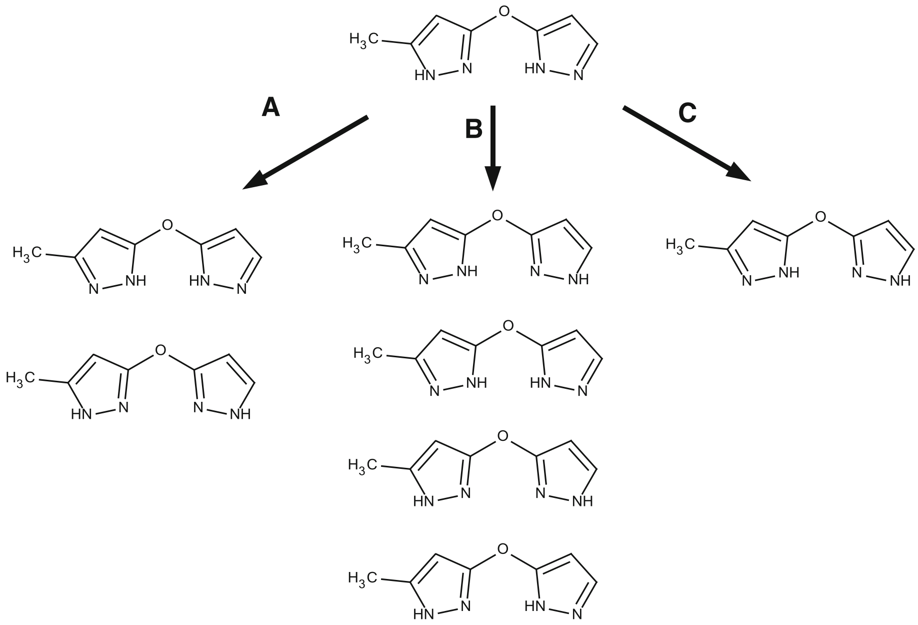


Fig. 1 Tautomer enumeration for pyrazoles with MUDO. Non-exhaustive enumeration (**a**) flips each pyrazole tautomer in turn and generates just two tautomers. Exhaustive enumeration (**b**) generates all four tautomers of the input structure. Generally, one should perform enumeration exhaustively for tautomers although the ability to enumerate forms non-exhaustively can be advantageous when identifying matched molecular pairs. The canonical tautomer is obtained by sorting the list of SMILES strings for the exhaustively enumerated tautomers strings and selecting only the tautomer corresponding to the first member of the list (**c**). The following SMIRKS was used for each of the tautomer enumerations: [H:1][n:2]1[nX2:3][c:4][c:5][c:6]1≫[n:2]1[n:3]([H:1])[c:4][c:5][c:6]1

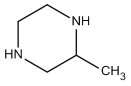
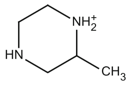
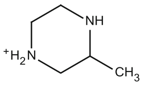
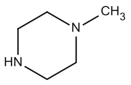
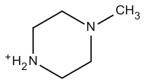
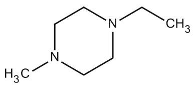
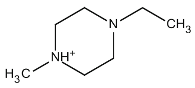
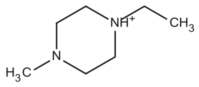
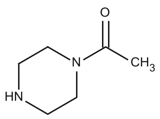
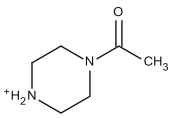
Neutral starting structures	Cationic forms generated in protonation step	Cationic forms generated in enumeration step
		
		
		
		

Fig. 2 Protonation of piperazines using MUDO is carried out in two steps. Firstly, a proton is added to one of the nitrogen atoms in the ring and this is achieved using the $NtoCat vector binding (Table 1) to prevent protonation of nitrogen that is linked by two carbons to cationic nitrogen. Separating the definitions for secondary and tertiary amine nitrogen and having secondary amines protonate before tertiary amines ensures that an *N*-alkylpiperazine will only be protonated on the unsubstituted nitrogen. In the subsequent enumeration step alternative forms are only generated when both ring nitrogen atoms are of the same type (secondary amine; tertiary amine)

of compounds can be seen in terms of relationships between chemical structures [1, 2, 29, 32] and the reader is directed to some of the work of Bajorath et al. [33, 34] who have published a number of important studies in this area. In this distance-geometric view of chemical space, compounds can also be described and characterized by their neighborhoods [35, 36]. Perhaps there is value in thinking of neighborhoods as having shape.

Table 1 Vector bindings for applying amine protonation model

Name	SMARTS definition
Csp3	[CX4]
SecAmine	[N;H1]([$Csp3])[$Csp3]
TerAmine	[N;H0]([$Csp3])([$Csp3])[$Csp3]
NtoCat	NCC[N+]
Prot1	[$SecAmine;!$NtoCat]
Prot2	[$TerAmine;!$NtoCat]

Results of MMPA can be used to guide molecular design [37–40]. For example, it was shown [29] that bioisosteric replacement of a carboxylic acid with tetrazole [41] is likely to lead to an increase in plasma protein binding even though the octanol/water logP [42] of acetic acid (−0.17) is greater than that of 5-methyltetrazole (−0.49). Knowledge of bioisosteric relationships [43–47] is greatly enhanced when linked to changes in measured physicochemical properties and ADMET characteristics that are associated with bioisosteric replacements. In a particularly thorough study, Boström et al. first show that 1,3,4-oxadiazoles typically have superior ADMET profiles when compared to the corresponding 1,2,4-oxadiazoles before relating these observations to electrostatic differences between the isomeric heteroaromatic rings [48]. Insights gained from MMPA of in vivo data are particularly valuable and a noteworthy example is provided by the observation by Sutherland, Watson et al. that replacing isopropyl bound to aromatic carbon with cyclopropyl leads to a mean reduction in unbound clearance of 0.37 log units [49]. This may reflect the relatively high carbon-hydrogen bond dissociation energy [50, 51] of cyclopropane that can be linked to strain in the three-membered ring. Another property of cyclopropyl relevant to molecular design is that it weakens bases to a greater extent than comparable substituents such as isopropyl [52].

In this article we use two examples to illustrate how molecule editing can be used to identify MMPs and in each case ChEMBL [53] was used as the source of data. Although software [54, 55] has been developed specifically for identification of MMPs, useful results can still be obtained using a general-purpose molecule editor. Furthermore, specialized MMPA software is not always able to capture the substructural context [29, 56] of the structural relationship. In the first example, MUDO was used to explore SAR for inhibitors of the cysteine protease inhibitor Cruzain that is of interest as a target [57] for intervention in the treatment of Chagas Disease [58]. The general structure of the inhibitors is shown in Fig. 3 and, since the compounds inhibit the enzyme irreversibly, activity is quantified as inactivation rate constant (k_{inact}) divided by inhibition constant (K_i). The compounds, which had all been described in a single article [59], were identified in ChEMBL by assay and by substructural search. MUDO was used to break the bond linking the S3 group to the rest of the molecule (Fig. 3) and the results of MMPA

Table 2 SMIRKS for applying amine protonation model

Description	SMIRKS
Protonate secondary amine in first step	[$Prot1:1]≫[$Prot1;h2;+:1]
Protonate tertiary amine in first step	[$Prot2:1]≫[$Prot2;h1;+:1]
Generate alternative protonation state in second step for mono-cation when both nitrogens are tertiary amines	[H:8][N+:1]1([C:7])[C:2][C:3][N;$TerAmine;+0:4][C:5][C:6]1≫[N+0:1]1([C:7])[C:2][C:3][N+:4]([H:8])[C:5][C:6]1
Generate alternative protonation state in second step for mono-cation when both nitrogens are secondary amines	[H:8][N+:1]1([H:7])[C:2][C:3][N;$SecAmine;+0:4]([H:9])[C:5][C:6]1≫[N+0:1]1([H:7])[C:2][C:3][N+:4]([H:8])([H:9])[C:5][C:6]1

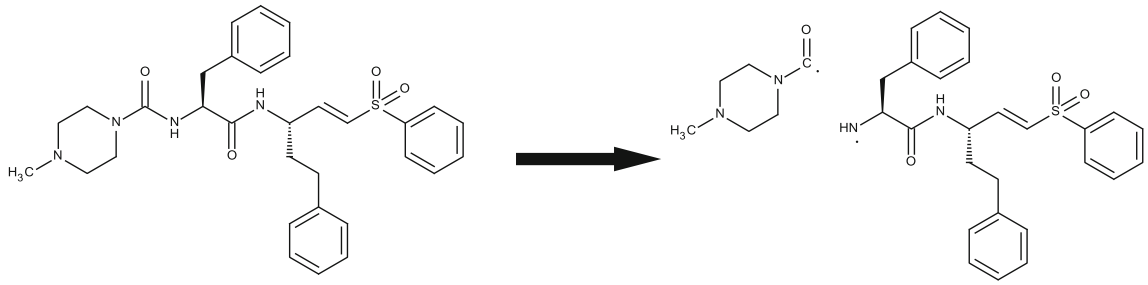
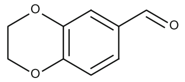
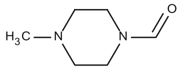
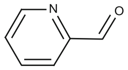
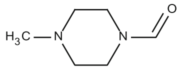
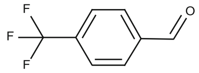
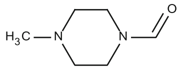
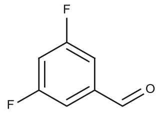
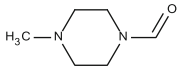
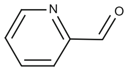
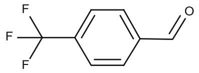
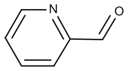
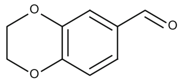
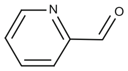
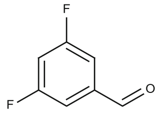
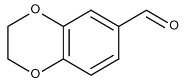
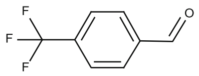
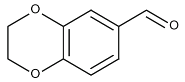
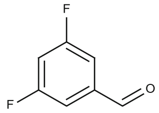
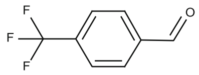
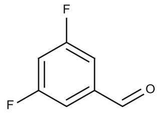


Fig. 3 Structural transformation used for MMPA of vinyl sulfone Cruzain inhibitors illustrated with K777. Each inhibitor was partitioned into its S3 group and the residual portion of the molecule by breaking the amide bond between the phenylalanine nitrogen and the carbonyl carbon of S3 substituent before converting to a pair of canonical SMILES strings. Each MMP shares a common residual portion of the molecule and can be identified by joining the SMILES pairs by this field

Table 3 Matched molecular pair analysis of Cruzain inhibition by vinyl sulfones

MMP	Group 1	Group 2	N [a]	Mean(ΔAct) [b,c]	Range(ΔAct) [b,d]
1			1	1.26	-
2			2	1.04	0.38
3			1	0.83	-
4			1	0.39	-
5			3	0.26	0.91
6			3	0.21	1.70
7			3	0.07	1.05
8			3	0.04	0.89
9			3	-0.15	1.84
10			3	-0.19	0.95

The data file associated with this analysis is included in the supplemental material

[a] Number of matched molecular pairs

[b] Act = log(k_{inact}/K_i) from ChEMBL assay 461748 [59]

[c] Mean difference ($Act_2 - Act_1$) in activity

[d] Range in difference in activity

are shown in Table 3. Fragmenting molecules in this manner allows multiple MMP relationships to be established simultaneously and this could also have also been achieved using Leatherface. The fragmented structures were written as disconnected SMILES strings which were split into scaffold and substituent SMILES. Matched molecular pairs were identified as compounds with the same scaffold SMILES and distinguished by their substituent SMILES. The analysis shows the 4-methylpiperazinylcarboxy S3 substituent to be associated with the highest levels of activity. Large variation in the activity difference observed for a particular MMP relationship can be interpreted as evidence for non-additivity in SAR and the largest ranges were observed for MMP relationships 6 (1.7 log units) and 9 (1.8 log units). The list of substituent pairs for a particular substitution position provides a richer description of chemical diversity within a set of analogs than does the list of the substituents and might be used to assess coverage of chemical space by different sets of analogs.

Relationships between chemical structures can be described as symmetrical or unsymmetrical according to whether one of defining substructures can be designated unambiguously as a reference. The relationship between enantiomers [60] is symmetrical as is that between pyrazole tautomers lacking substituents on nitrogen and the tautomerism can also be described as degenerate [1, 2]. In contrast, the relationship between 2-hydroxypyridine and its 2-pyridone tautomer is unsymmetrical because either tautomeric form could be used as a reference. Average differences are typically of no interest when performing

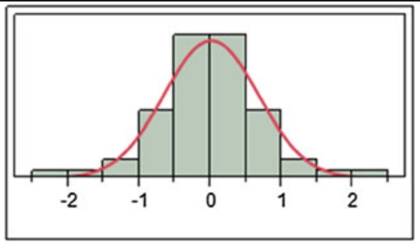
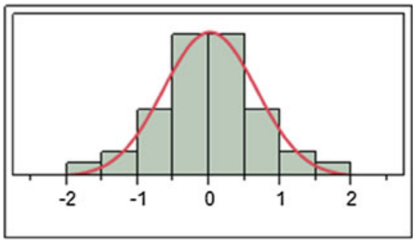
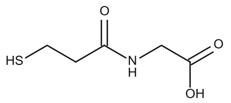
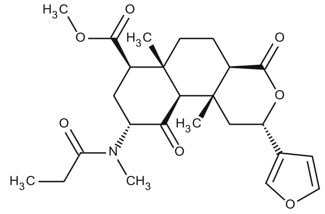
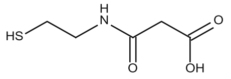
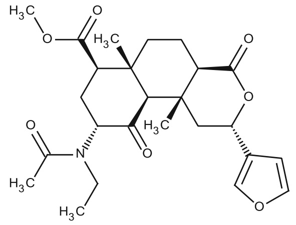
MMPA for symmetrical relationships between structures and the focus of analysis is variation in property value differences.

In a second example of MMPA we investigated the effect of ‘amide reversal’, which can be regarded as a bioisoteric replacement (albeit a ‘degenerate’ one), on potency as shown in Fig. 4. Amides were only reversed if the carbon atoms bonded to the amide core were achiral and this can be specified by requiring that the relevant carbon atoms have at least two hydrogen atoms attached. Analysis was performed separately for secondary and tertiary amides and, in some cases, two new structures were generated when the groups on tertiary nitrogen were different. The MMPs were identified as matching SMILES for the original and edited structures and each is found twice (in opposite orders) because the structural relationship defining the MMP is symmetrical. Although it is possible to select just one instance of each MMP and use the absolute value of the activity difference, retaining both instances makes the symmetry in the relationship explicit and allows measures of variation to be interpreted in the usual manner. The results of the analysis are shown in Fig. 4 with the MMPs showing the greatest differences. The standard deviations of the secondary amide (0.66) and tertiary amide (0.65) distributions are essentially identical. The inter-quartile ranges for the secondary (0.79) and tertiary (0.59) amides also suggest that the average effects of amide reversal on affinity/potency are relatively small and amide reversal should be considered a conservative structural modification in the context optimizing affinity or potency. Reversed amides would not in general be identified by MMPA algorithms based on maximal common substructure.

Modeling covalently-bound ligands

Formation of covalent bonds between ligands and their target proteins can be exploited to enhance both affinity and selectivity of interactions. Creation of covalent bonds between ligand and protein in docking presents challenges because the covalent geometry of each must be modified in

Fig. 4 MMPA of reversal of secondary (**a**) and tertiary (**b**) amides showing SMIRKS used to carry out each transformation. Carbon atoms bonded to amide groups were required to be saturated with two or three hydrogen atoms. Also shown are the activity ratio distributions for are the pair for which the largest difference in activity was observed. Activity was quantified by IC_{50}, K_i or K_d and only in-range data was used in the analysis and each pair of activities were required to be of the same type (e.g. IC_{50}) with the same ChEMBL assay code. Data analysis was performed using JMP (http://www.jmp.com) and the data sets are included in the supplementary information

A	B
SMIRKS ([$C] is sp^3 carbon with 2 or 3 hydrogens): [$C:1][C:2](=[O:3])[N:4]([H:5])[$C:6]>> [C:6][C:2](=[O:3])[N:4]([H:5])[C:1]	SMIRKS ([$C] is sp^3 carbon with 2 or 3 hydrogens): [$C:1][C:2](=[O:3])[N:4]([CX4:5])[$C:6]>> [C:6][C:2](=[O:3])[N:4]([C:5])[C:1]
log(activity$_1$/activity$_2$)	log(activity$_1$/activity$_2$)
	
	 H3C O O H3C O O O H3C O N H3C CH3 O
IC_{50} = 2.8 µM	K_i = 7 nM
	 H3C O O H3C O O O H3C N H3C H3C O
IC_{50} = 580 µM	K_i = 280 nM

order to form the bond. However, the more stringent geometric constraints associated with covalent bonds also facilitate prediction of binding modes and docking software that has not been explicitly designed for covalently-bound complexes can be adapted for this purpose. A molecule editor can be used both to modify ligands prior to docking and then to link the docked ligands covalently to the protein. Once protein and ligand have been linked covalently, the model of the complex can be refined using empirical force field methods.

We use the example of the vinyl sulfone inhibitor K777 bound to Cruzain [61] to illustrate the role of MUDO in building models for covalently bound complexes. Firstly, MUDO was used to add hydrogen sulfide across the carbon–carbon double bond of K777 so that the ligand has the appropriate covalent structure. The configuration of the chiral center created by hydrogen sulfide addition was unspecified in the SMILES representation of the reaction product. OMEGA [62, 63] was used to build a single conformation corresponding to each configuration of the carbon to which the thiol is bonded and the two stereoisomers can be labeled SSS and RSS (configurations of the other two chiral centers are fixed). Both stereoisomers were energy-minimized using the MMFF94S [64] force field and Poisson-Boltzmann implicit water model as implemented in SZYBKI [65]. The protein model used for the docking consisted of the A-chain protein from the Cruzain-K777 complex [61] in which the catalytic cysteine (C25) has been pruned to glycine in order to avoid clashes between the protein and ligand thiol. The two stereoisomeric ligands were docked using Glide [66] with a restraint imposed on the ligand thiol to position it near where the thiol would be in the unmodified protein. The best score (−9.33) observed for the SSS stereoisomer (bound configuration in crystal structure) was more favorable than the corresponding value (−7.31) for the RSS stereoisomer. MUDO was then used to link the ligand to Cruzain itself (the ligand thiol was deleted in the linking process) and the resulting structure was energy-minimized. A comparison of the crystallographic and modeled ligand structures is shown in Fig. 5.

Summary

We have illustrated some of the diverse ways in which a general-purpose molecule editor can be used in molecular design and hope to have shown how line notations like SMIRKS and SMARTS can be used to impose views of chemistry on large numbers of molecular structures in an objective and transparent manner. We have also made source code available as supplemental material and hope this will help other researchers to find new applications for automated molecule editing.

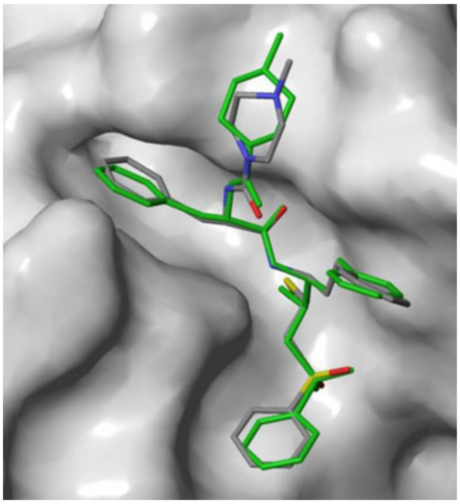


Fig. 5 Comparison of crystallographic binding mode of K777 with modeled structure (*green*). After the docked ligand had been linked to Cruzain using MUDO, the covalently-bound complex was energy minimized (OPLS_2005; continuum solvation; protein backbone carbon and nitrogen coordinates frozen) using the Impact application within Maestro (http://www.schrodinger.com/productpage/14/12/)

Acknowledgments We are grateful to the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; Grant Numbers #2011/01893-3 and #2011/20572-3)) and the Conselho Nacional de Pesquisa (CNPq) for financial support. We thank OpenEye for an academic software license and the two anomymous reviewers for their helpful and constructive comments.

References

1. Kenny PW, Sadowski J (2005) Structure modification in chemical databases. Methods and principles in medicinal chemistry. In: Oprea T (ed) Chemoinformatics in drug discovery. 23:271–285
2. Barnard JM, Kenny PW, Wallace PN (2012) Representing chemical structures in databases for drug design. RSC drug discovery series 13 (drug design strategies) 164–191
3. Southan C, Várkonyi P, Muresan S (2007) Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics. Curr Top Med Chem 7:1502–1508
4. Lyne PD, Kenny PW, Cosgrove DA, Deng C, Zabludoff S, Wendoloski JJ, Ashwell S (2004) Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening. J Med Chem 47:1962–1968
5. Krumrine JR, Maynard AT, Lerman CL (2005) Statistical tools for virtual screening. J Med Chem 48:7477–7481
6. Irwin JJ, Shoichet BK (2005) ZINC-a free database of commercially available compounds for virtual screening. J Chem Inf Sci 45:177–182
7. Walters WP, Stahl MT, Murcko MA (1998) Virtual screening–an overview. Drug Discov Today 3:160–178
8. SMIRKS Theory Manual, Daylight Chemical Information Systems, Inc., Laguna Niguel, CA 92677. http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html. Accessed 16 Dec 2012
9. Daylight toolkit, Daylight Chemical Information Systems, Inc., Laguna Niguel, CA 92677. http://www.daylight.com/products/toolkit.html. Accessed 16 Dec 2012

10. SMARTS Theory Manual, Daylight Chemical Information Systems, Inc., Laguna Niguel, CA 92677. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed 20 May 2013
11. Bartlett PA, Shea GT, Telfer SJ, Waterman S (1989) CAVEAT: a program to facilitate the structure-derived design of biologically active molecules. Special publication-Royal Society of Chemistry. Mol Recognit Chem Biochem. Probl 78, 182–196
12. Schneider G, Neidhart W, Giller T, Schmid G (1999) "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening. Angew Chem Int Ed 38:2894–2896
13. Unity. Tripos International, St. Louis, MO 63144-2319. http://www.tripos.com/index.php?family=modules,SimplePage,,,&page=UNITY. Accessed 25 May 2013
14. Van Drie JH, Weininger D, Martin YC (1989) ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. J Comput Aided Mol Des 3:225–251
15. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comp Sci 28:31–36
16. Weininger D, Weininger A, Weininger JL (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comp Sci 29:97–101
17. OEChem Toolkit Manual, OpenEye Scientific Software, Santa Fe, NM 87508. http://www.eyesopen.com/docs/toolkits/current/html/OEChem_TK-c++/index.html. Accessed 26 Oct 2012
18. Morley AD, Kenny PW, Burton B, Heald RA, MacFaul PA, Mullett J, Page K, Porres SS, Ribeiro LR, Smith P, Ward S, Wilkinson TJ (2009) 5-Aminopyrimidin-2-ylnitriles as cathepsin K inhibitors. Bioorg Med Chem Lett 19:1568–1661
19. Elguero J, Marzin C, Katritzky AR, Lind P (1975) The tautomerism of heterocycles. Adv Heterocycl Chem Suppl 1:1–656
20. Button RG, Cairns JP, Taylor PJ (1985) Tautomeric ratio in 4-methylthiazol-2-ylguanidine, a model guanidinoheterocycle. J Chem Soc Perkin Trans 2:1555–1558
21. Albert A, Taylor PJ (1989) The tautomerism of 1,2,3-triazole in aqueous solution. J Chem Soc Perkin Trans 2:1903–1905
22. Martin YC (2009) Let's not forget tautomers. J Comput Aid Mol Des 23:693–704
23. Sayle RA (2010) So you think you understand tautomerism? J Comput Aid Mol Des 24:485–496
24. Oellien F, Cramer J, Beyer C, Ihlenfeldt W-D, Selzer PM (2006) The impact of tautomer forms on pharmacophore-based virtual screening. J Chem Inf Model 46:2342–2354
25. Claramunt RM, Garcia MA, Lopez C, Trofimenko S, Yap GPA, Alkorta I, Elguero J (2005) The tautomerism of 1H-pyrazole-3(5)-(*N*-tert-butyl)carboxamide in the solid state and in solution. Magn Reson Chem 43:89–91
26. Khalili F, Henni A, East ALL (2009) pK_a values of some piperazines at (298, 303, 313, and 323) K. J Chem Eng Data 54:2914–2917
27. Hammett LP (1937) Effect of structure upon the reactions of organic compounds Benzene derivatives. J Am Chem Soc 59: 96–103
28. Free SM, Wilson JW (1964) A mathematical contribution to structure-activity studies. J Med Chem 7:395–399
29. Birch AM, Kenny PW, Simpson I, Whittamore PRO (2009) Matched molecular pair analysis of activity and properties of glycogen phosphorylase inhibitors. Bioorg Med Chem Lett 19:850–853
30. Zwanzig RW (1954) High-temperature equation of state by a perturbation method I. Nonpolar gases. J Chem Phys 22:1420–1426
31. Shirts MR, Mobley DL, Chodera JD (2007) Alchemical free energy calculations: ready for prime time? Ann Rep Comp Chem 3:41–59
32. Maggiora GM (2006) On outliers and activity cliffs-why QSAR often disappoints. J Chem Inf Model 46:1535
33. Wassermann AM, Wawer M, Bajorath J (2010) Activity landscape representations for structure—activity relationship analysis. J Med Chem 53:8209–8223
34. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry. J Med Chem 55:2932–2942
35. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE (1996) Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. J Med Chem 39:3049–3059
36. Blomberg N, Cosgrove DA, Kenny PW, Kolmodin K (2009) Design of compound libraries for fragment screening. J Comput Aid Mol Des 23:513–525
37. Leach AG, Jones HD, Cosgrove DA, Kenny PW, Ruston L, MacFaul P, Wood JM, Colclough N, Law B (2006) Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. J Med Chem 49:6672–6682
38. Griffen E, Leach AG, Robb GR, Warner DJ (2011) Matched molecular pairs as a medicinal chemistry tool. J Med Chem 54:7739–7750
39. Wassermann AM, Dimova D, Iyer P, Bajorath J (2012) Advances in computational medicinal chemistry: matched molecular pair analysis. Drug Dev Res 73:518–527
40. Dossetter AG, Griffen EJ, Leach AG (2013) Matched molecular pair analysis in drug discovery. Drug Discov Today 18:724–731
41. Herr RJ (2002) 5-Substituted 1H-tetrazoles as carboxylic acid isosteres: medicinal chemistry and synthetic methods. Bioorg Med Chem 10:3379–3393
42. LOGKOW, A databank of evaluated octanol-water partition coefficients: http://logkow.cisti.nrc.ca/logkow/index.jsp. Accessed 26 Oct 2012
43. Thornber CW (1979) Isosterism and molecular modification in drug design. Chem Soc Rev 8:563–580
44. Patani GA, LaVoie EJ (1996) Bioisosterism: a rational approach in drug design. Chem Rev 96:3147–3176
45. Sheridan RP (2002) The most common chemical replacements in drug-like compounds. J Chem Inf Comp Sci 42:103–108
46. Meanwell Nicholas A (2011) Synopsis of some recent tactical application of bioisosteres in drug design. J Med Chem 54:2529–2591
47. Papadatos G, Brown N (2013) In silico applications of bioisosterism in contemporary medicinal chemistry practice. WIREs Comput Mol Sci 3:339–354
48. Boström J, Hogner A, Llinas A, Wellner E, Plowright AT (2012) Oxadiazoles in medicinal chemistry. J Med Chem 55:1817–1830
49. Sutherland JJ, Raymond JW, Stevens JL, Baker TK, Watson DE (2012) Relating molecular properties and in vitro assay results to in vivo drug disposition and toxicity outcomes. J Med Chem 55:6455–6466
50. Bach RD, Dmitrenko O (2004) Strain energy of small ring hydrocarbons. Influence of C–H bond dissociation energies. J Am Chem Soc 126:4444–4452
51. Tian Z, Fattahi A, Lis L, Kass SR (2006) Cycloalkane and cycloalkene C–H bond dissociation energies. J Am Chem Soc 128:17087–17092
52. Perrin CL, Fabian MA, Rivero IA (1999) Basicities of cycloalkylamines: Baeyer strain theory revisited. Tetrahedron 55:5773–5780
53. ChEMBL version 15. http://www.ebi.ac.uk/chembl. Accessed 30 May 2013
54. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50:339–348

55. Warner DJ, Griffen EJ, St-Gallay SA (2010) WizePairZ: a novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. J Chem Inf Model 50:1350–1357
56. Papadatos G, Alkarouri M, Gillet VJ, Willett P, Kadirkamanathan V, Luscombe CN, Bravi G, Richmond NJ, Pickett SD, Hussain J, Pritchard JM, Cooper AWJ, MacDonald SJF (2010) Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of hERG inhibition, solubility, and lipophilicity. J Chem Inf Model 50:1872–1886
57. Engel JC, Doyle PS, Hsieh I, McKerrow JH (1998) Cysteine protease inhibitors cure an experimental Trypanosoma cruzi infection. J Exp Med 188:725–734
58. Chagas C (1909) Nova tripanozomiaze humana: estudos sobre a morfolojia e o ciclo evolutivo do Schizotrypanum Cruzi n. gen., n. sp., ajente etiolojico de nova entidade morbida do homem. Mem Inst Oswaldo Cruz 1:159–218
59. Jaishankar P, Hansell E, Zhao D-M, Doyle PS, McKerrow JH, Renslo AR (2008) Potency and selectivity of P2/P3-modified inhibitors of cysteine proteases from trypanosomes. Bioorg Med Chem Lett 18:624–628
60. Leach AG, Pilling EA, Rabow AA, Tomasi S, Asaad N, Buurma NJ, Ballard A, Narduolo S (2012) Enantiomeric pairs reveal that key medicinal chemistry parameters vary more than simple physical property based models can explain. Med Chem Commun 3:528–540
61. Kerr ID, Lee JH, Farady CJ, Marion R, Rickert M, Sajid M, Pandey KC, Caffrey CR, Legac J, Hansell E, McKerrow JH, Craik CS, Rosenthal PJ, Brinen LS (2009) Vinyl sulfones as antiparasitic agents and a structural basis for drug design. J Biol Chem 284:25697–25703
62. OMEGA.OpenEye Scientific Software, Santa Fe, NM 87508. http://www.eyesopen.com/omega. Accessed 28 Feb 2013
63. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) Conformer Generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and Cambridge structural database. J Chem Inf Model 50: 572–584
64. Halgren TA (1999) MMFF VI. MMFF94S option for energy minimization studies. J Comp Chem 20:720–729
65. SZYBKI. OpenEye Scientific Software, Santa Fe, NM 87508. http://www.eyesopen.com/szybki. Accessed 28 Feb 2013
66. GLIDE. http://www.schrodinger.com/productpage/14/5/21/. Accessed 30 May 2013