

Division of Pharmaceutical Chemistry and Technology
Faculty of Pharmacy
University of Helsinki
Finland

Investigating phosphate structural replacements through computational and experimental approaches

by

Yuezhou Zhang

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Pharmacy of the University of Helsinki, for public examination in Walter Hall (2089) at the EE building (Agnes Sjöbergin katu 2), on December 19th 2014, at 12:00 noon.

Helsinki 2014

Supervisors	Dr. Henri Xhaard Division of Pharmaceutical Chemistry and Technology Faculty of Pharmacy, University of Helsinki, Finland
	Professor Jari Yli-Kauhaluoma Division of Pharmaceutical Chemistry and Technology Faculty of Pharmacy, University of Helsinki, Finland
	Dr. Gustav Boije af Gennäs Division of Pharmaceutical Chemistry and Technology Faculty of Pharmacy, University of Helsinki, Finland
Reviewers	Docent Konstantin Denessiouk Department of Biosciences, Åbo Akademi University, Finland Graduate School of Engineering, Tohoku University, Japan
	Docent Tiina Salminen Department of Biosciences, Åbo Akademi University Finland
Opponent	Professor Jonathan B. Baell Department of Medicinal Chemistry Faculty of Pharmacy and Pharmaceutical Sciences Monash University Australia

© Yuezhou Zhang 2014
 ISBN 978-951-51-0044-3 (Paperback)
 ISBN 978-951-51-0045-0 (PDF)
 ISSN 2342-3161

Reproduced in part with permission from the ACS Medicinal Chemistry Letters (submitted manuscript) and the Journal of Chemical Information and Modeling (to be submitted for publication). Material (figures and text) from sections 4.2, 4.3, and 5.5. Zhang, Y., Ghemtio, L., Venkannagari, H., Lehtiö, L., Yli-Kauhaluoma, J., Xhaard, H., Boije af Gennäs, G. Evaluation of adenosine analogs bearing phosphate isosteres at the human macromolecule-containing protein 1. Material (figures and text) from sections 4.1, and 5.1-5.4. Zhang, Y., Borrel, A., Boije af Gennäs, G., Yli-Kauhaluoma, J., Xhaard, H. Phosphate local structural isosteric replacements extracted from the Protein Data Bank. Unpublished work copyright 2014 American Chemical Society.

Helsinki University Printing House
 Helsinki 2014

Abstract

Bioisosteric replacements are used in drug design during lead generation and optimization processes with the aim to replace one functional group of a known molecule by another while retaining biological activity. The reason to use bioisosteric replacements are typically to optimize bioavailability or reducing toxicity. Phosphate groups represent a paradigm to study bioisosteric replacements. Protein-phosphate interaction plays a critical role during molecular recognition processes, and for example kinases represent one of the largest families of drug targets. However, some challenges exclude phosphate as a promising lead-like building block: i) charged phosphates do not cross molecular membranes; ii) some widely expressed proteins such as phosphatases easily hydrolyze phosphoric acid esters, which lead phosphate-containing ligands to lose their binding affinities before reaching their biological targets; iii) introduction of phosphate groups to parent scaffold is not easy.

In the first part of the thesis work, I designed and implemented a computational protocol to mine information about phosphate structural replacements deposited in the Protein Data Bank. I constructed 116, 314, 271, and 42 sets of superimposed proteins where each set contains a reference protein to either POP, AMP, ADP, or ATP as well as a certain number of non-nucleotide ligands. 929 of such ligands are under study. The chemotypes that came out as structural replacements are diverse, ranging from common phosphate isosteres such as carboxyl, amide and squaramide to more surprising moieties such as benzoxaborole and aromatic ring systems. I exemplified some novel examples and interpreted the mechanism behind them. Local structural replacements are circumstance dependent: one chemical group valid in certain set-up cannot necessarily guarantee the success of another. The data from the study is available at http://86.50.168.121/phosphates_LSR.php.

In the second part, I synthesized fifteen compounds retaining the adenosine moieties and bearing bioisosteric replacements of the phosphate at the ribose 5'-oxygen to test their stability toward human macro domain protein 1. These compounds are composed of either a squaryldiamide or an amide group as the bioisosteric replacement and/or as a linker. To these groups a variety of substituents were attached: phenyl, benzyl, pyridyl, carboxyl, hydroxy and tetrazolyl. Biological evaluation using differential scanning fluorimetry showed that four compounds stabilized human MDO1 at levels comparable to ADP and one at level comparable to AMP. Virtual screening was also run to identify MDO1 binding ligands. Among 20,000 FIMM database lead-like molecules, 39 compounds were selected for testing and eleven compounds found active based on ADPr and Poly-ADPr competition binding assay. The assay is however not well validated and a second confirmatory assay was conducted using calorimetry. To the best of my knowledge, this is the first report of non-endogenous ligands of the human MDO1.

Altogether, this thesis highlights the versatility of molecular recognition processes that accompanies chemical replacements in compounds; this in turns shows the limits of the concepts of molecular similarity and “classical” bioisosterism that are based on the conservation of molecular interactions.

Acknowledgements

This thesis work was carried out at the Division of Pharmaceutical Chemistry and Technology, Faculty of Pharmacy, University of Helsinki from December 2009 to December 2014. I am highly grateful for the financial support granted by China Scholarship Council and University of Helsinki. The Finnish National Doctoral Program in Informational and Structural Biology is thanked for organizing graduate studies and providing support to graduate education.

First of all, I would like to express my deepest gratitude towards my main supervisor, Dr. Henri Xhaard for giving me a chance to work. His patience, guidance, and encouragement he provided are pillars to drive my doctoral studies. His ethic working spirit has dramatically impressed me, which will stamp into my heart as a real researcher. Henri's care to me can be metaphorized as a baby-sitter taking care of a new-born baby. What he tried to help me is comprehensive: from teaching the English in the beginning to scientific thinking until making a long-term career plan. His explorative philosophy to scientific research always prompts me to try different things, which not only broaden the horizon to specific topics, but keep the studies attractive. Henri's special management skills in our international research group make me always feel at home even I am far away from my family.

I am very happy to have a supervisor team to get along with. Besides Henri, my co-supervisors Professor Jari Yli-Kauhaluoma and Dr. Gustav Boije af Gennäs are all with me whenever I need help. Jari is such kind person who always delivers positive energy to fuel my research work. His broad and profound knowledge make me felt so lucky to work with him. Whenever I have a question or need his advice, he can answer me fast with great enthusiasm though he is very busy. Gusse is not only my supervisor but best Finnish friend of mine so far. He explains me scientific problem very clearly and patiently. I enjoy it! We also play football together during summer time. A lot exited moment when we are together is in my memory. I have also experienced several exciting moments with him which I will not forget. With that, I owe all my supervisors big thanks.

For reviewing this thesis on a super busy schedule, Docent Tiina Salminen and Docent Konstantin Denessiouk are greatly acknowledged for their valuable comments which have been incorporated into this special thesis.

I also would like to thank all of my collaborators and co-authors, especially Alexandre Borrel and Harikanth Venkannagari for their contribution to this work. Alexandre teaches me so many script skills which is the key to my project. Hari performed the biological wet lab experiments on time though he has his own projects in hand. Docent Lari Lehtiö's constructive comments on manuscript are essential.

Many thanks to my colleagues, Leo Ghemtio, Gloria Wissel, Michal Stepniewski, Ainoleena Turku, Lasse Karhu, Maiju Rinne, Vigneshwari Subramanian, Aniket Magarkar together with two new Master students. The research is challenging, but we still organize group activities to release the press. For me, Leo is amazing person. He is able to manage so many things at the same time. Leo is the IT supporter of our group, including me. Gloria is innate organizer and manager; she can always bring us happiness and surprise. I fetched many help from her such as she helped me organize my defense. Michal is very

generous to share knowledge with me. Ainoleena is the first groupmate I met when went to our office and obviously I got first help from her. Lasse and Aniket are a so smart therefore I can ask them so many questions. Vignishwari's distinct career plan enlightens me think about myself. Maiju is pinpoint compass when we enjoy cross-country skiing. I do not have many talks with two new students, but I felt they very easy-going people. I also want to say thank you to all of my colleagues at the synthesis group and among the group I especially want to acknowledge Ingo Aumüller and Mikko Vahermo for the friendship.

Apart from wonderful research group, I have been blessed with an excellent group of friends. Taking this opportunity I thank them: Dongfei Liu for great discussions and cooperation; Hongbo Zhang's guide me to start up my life in Finland; Zuyue Chen has been with me to play badminton and encouraged me a lot; Liwei Liu's promotion when I was depressed; Zhilin Li, Li Ma, Yan Yan together having excited communication; Changfang Wang's help.

In addition, I would like to dedicate this dissertation to my family, especially my wife, Juan. This research will not finish without her encouragement and understanding. Of course, the support from my parents is one of corner-stones of my PhD journey. Thank you all very much!

Final, I would like to thank everybody who was for not only the research but the daily life, as well as express my apology if I missed on some. The life at this stage may draw a period, but it is the start point of next stage. I will conserve all these cherished memories and friendship when I embark the next phase of my life. I am also eager to keep contact with all of you.

Contents

Abstract	i
Acknowledgements	ii
Contents.....	iv
Abbreviations	1
1 Introduction	2
2 Review of the literature.....	4
2.1 Molecular recognition	4
2.1.1 Ligand-receptor interactions.....	5
2.1.1.1 Thermodynamics of ligand binding.....	5
2.1.1.2 Solvent effects in protein-ligand recognition	7
2.1.2 Specific intermolecular interactions (noncovalent interactions)	9
2.1.2.1 Hydrogen bonds.....	11
2.1.2.2 Halogen bonds	12
2.1.2.3 π interactions	14
2.2 Chemical replacement strategies	15
2.2.1 The Structure-Property Principle.....	15
2.2.1.1 Defining molecular similarity.....	15
2.2.1.2 Activity cliffs.....	16
2.2.2 Isosteres and bioisosteres	16
2.2.2.1 The development of bioisosterism.....	16
2.2.2.2 Phosphate isosteres	17
Phosphorus-containing isosteres	17
Carboxylic acid.....	19
Sulfur-containing isosteres	20
Boronic acids isosteres	20
Squaramide-based isosteres.....	21
Other acidic replacements of phosphate.....	22
2.2.3 Scaffold hopping	24
2.2.3.1 Heterocyclic replacements.....	24
2.2.3.2 Ring closure and opening	25
2.2.3.3 Peptide mimetics	27
2.2.4 Other considerations	28
2.2.4.1 Isosteric replacements are target-dependent.....	28
2.2.4.2 Prodrugs to deliver isosteres.....	29
2.3 Molecular similarity in a chemoinformatics perspective	30
2.3.1 Databases to study isosteric replacement	30
2.3.2 Ligand-based drug discovery: analog search and QSAR	31
2.3.2.1 Representing molecules to measure their similarity.....	31
2.3.2.2 Adjacency/connectivity matrices and graph-based representation.....	32
2.3.2.3 Numerical molecular descriptors.....	33

2.3.2.4 Molecular fingerprints	34
2.3.2.5 Shape and electrostatics.....	35
2.3.2.6 Pharmacophoric representation	35
3 Aims of the study	37
4 Materials and Methods	38
4.1 Computational section	38
4.1.1 Material.....	38
4.1.2 Methods	39
4.1.3 Homology modeling, docking and virtual screening	45
4.2 Synthetic chemistry section	47
4.3 Biological testing.....	48
5 Results and discussion.....	52
5.1 Workflow to mine the Protein Data Bank for local structural replacements	53
5.2 Previously unrecognized structural replacements of phosphate.....	64
5.3 Structural replacement of ribose.....	69
5.4 Molecular mechanisms involved in the recognition of local structural replacements	73
5.5 Discovery of compounds binding to MDO1	77
5.5.1 Modeling MDO1 – structural analysis	77
5.5.2 Synthesis of ADP analogs	79
5.5.2.1 Chemistry	80
5.5.2.2 Biological activity	83
5.5.2.3 Docking analysis	87
5.6 Virtual screening approach.....	90
6 Perspectives	96
References	100

Abbreviations

3D	Three-dimensional	DSF	Differential scanning fluorimetry
AMP	Adenosine-5'-monophosphate	HIV	Human immunodeficiency virus
ADP	Adenosine-5'-diphosphate	ITC	Isothermal titration calorimetry
ATP	Adenosine-5'-triphosphate	LSR	Local structural replacement
ADPr	Adenosine-5'-diphosphate ribose	MDO1	Macro domain protein 1
Boc	<i>tert</i> -Butyloxycarbonyl	MDO2	Macro domain protein 2
BS	Binding site	HSP	Heat-shock protein
CDK2	Cyclin-dependent kinase 2	Pi1	α -phosphate
CSD	Cambridge Structure Database	Pi2	β -phosphate
DIPEA	<i>N,N</i> -Diisopropylethylamine	Pi3	γ -phosphate
HBTU	<i>N,N,N',N'</i> -Tetramethyl- <i>O</i> -(1 <i>H</i> -benzotriazol-1-yl)uronium hexafluorophosphate	PDB	Protein Data Bank
DMSO	Dimethyl sulfoxide	POP	Pyrophosphate
DSC	Differential scanning calorimetry	QSAR	Quantitative structure-activity relationship
		Å	Ångström, 10^{-10} m

Amino acid	3-Letter Symbol	1-Letter Symbol	Amino acid	3-Letter Symbol	1-Letter Symbol
Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartic acid	Asp	D	Proline	Pro	P
Cysteine	Cys	C	Phenylalanine	Phe	F
Glutamic acid	Glu	E	Serine	Ser	S
Glutamine	Gln	Q	Threonine	Thr	T
Glycine	Gly	G	Tryptophan	Trp	W
Histidine	His	H	Tyrosine	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V

1 Introduction

Medicinal chemists face challenges in developing small molecules with high potency as well as other characteristics such as the capability to be delivered to biological targets (bioavailability), reduced toxicity, unwanted effects and increase the permeability (George & Nathan, 2013), together with chemical and metabolic stability and solubility. Often, when a compound is available, it needs to be modified and optimized for one or the other of these objectives in a process named lead optimization. Bioisosteric replacement is one of the main strategies to perform optimization. Bioisosteric replacements are fulfilled by replacing one functional group of a known molecule by other functional group resulting in a similar biological/physicochemical property. As stated by the IUPAC, “A bioisostere is a compound resulting from the exchange of an atom or group of atoms with another, broadly similar, atom or group of atoms.” (<http://goldbook.iupac.org/BT06798.html>)

Among the variety of chemical moieties that have been studied by medicinal chemists over the last century, phosphate groups are one of the most urgently demanded since there is a considerable amount of proteins whose endogenous ligand contains phosphate groups. For example, the single protein family of kinases accounts for about 1.7% of the proteins encoded by the human genome (Manning, et al., 2002). In medicinal chemistry, phosphate groups need to be replaced for several reasons: (1) the diffusion of negative charged phosphates across the highly hydrophobic cell membrane is highly unfavorable (Smith, et al., 2003); (2) some enzymes, phosphatases for example, hydrolyze phosphoric acid ester very easily, which probably let phosphate containing ligands to lose the majority of binding affinities without approaching their biological targets; (3) introduction of phosphate groups into parent scaffold or compounds is challenging in terms of chemistry (Zhao, et al., 2012).

In this thesis, taking the example of phosphate groups in particular, as well as 1',5'-bonded ribose, we investigated how they can be replaced by other chemotypes. We constructed a fully automated workflow to extract protein ligand structural information deposited in Protein Data Bank (PDB) (Berman, 2000) to automatically extract local structural replacements of the phosphate and ribose groups that were found in AMP, ADP, ATP and POP binding proteins. This data allows us to consider the different mechanisms

as well as replacement possibilities for phosphate and ribose. While this data does not take into account biological activity, it does provide key information about the mechanisms beyond the structural replacements. The data is made available to the research community through a website located at <http://86.50.168.121/bioisosteres.php>.

Secondly, phosphate (bio)isosterim is analyzed directly by synthesizing a small library of 15 ADPr analogs, belonging to two series that bear classical bioisosteric replacements in their phosphate moieties and were tested towards an human ADP-ribose (ADPr) binding protein, macro domain protein 1. The MDO1 has to date no synthetic compounds that could be used for biochemical studies or to be used as a template to construct a fluorescent biochemical probe for screening or molecular imaging. Due to the absence of such compounds, the functions of human MDO1 are still poorly understood. Recent findings have implicated human MDO1 has been implicated in multiple cancers, such as ovarian carcinoma (Tian, et al., 2009), breast carcinoma (Zhao, et al., 2010), colorectal carcinoma (Han et al., 2011), and acute leukemia (Imagama, et al., 2007). The MDO1 is able to hydrolyze *O*-acetyl-ADP-ribose and four crystal structures exists that show a complex of the unliganded human MDO1 (PDB code: 2X47; Chen et al., 2011) as well as the related parasitic MDO1 (PDB code: 2BFQ, 2BFR), (Karras, et al., 2005) and human MDO2 (PDB code: 4IQY),(Jankevicius, et al., 2013) these latter being bound with ADPr. Five compounds among fifteen were successfully found to stabilize MDO1 in millimolar concentrations using differential scanning fluorimetry, which relates to a binding constant in the μM range.

Finally, I present results about virtual screening of the human MDO1 protein. Forty-three ligands were screened using competitive assay and isothermal titration calorimetry (ITC) and differential scanning calorimetry (DSC). Nine compounds were found active using competition assay, however by an assay not fully validated. DSC was used to retest the compounds but at a concentration probably too low for detection. As a perspective, these experiments need to be conducted again.

2 Review of the literature

This literature review is based on three sections: 1) fundamental concepts about molecular recognition; 2) the concept of bioisosteric replacements including phosphate bioisosteres; 3) the concept of molecular similarity in a computational perspective. A number of excellent reviews have been used as a source to initiate the search of articles in these fields. For molecular interactions, I am referring to perspective article “*A medicinal chemist's guide to molecular interactions*” (Bissantz, et al., 2010). For phosphate bioisosterism, there are three reviews: “*Phosphate isosteres in medicinal chemistry*” (Rye and Baell, 2005), “*The use of phosphate bioisosteres in medicinal chemistry and chemical biology*” (Elliott et al., 2012) and “*Synthesis of α -brominated phosphonates and their application as phosphate bioisosteres*” (Downey and Cairo, 2014). For general bioisostere applications in drug design, a short course-based article “*Synopsis of Some Recent Tactical Applications of Bioisosteres in Drug Design*” was cited (Meanwell, 2011). The textbook “*Bioisosteres in medicinal chemistry*” from Nathan Brown was also a knowledgeable source of inspiration.

2.1 Molecular recognition

It is the recognition of ligands by their host molecules, typically proteins, which drives biological processes. Therefore, understanding this process is crucial to life sciences. Furthermore, molecular recognition is critical to structure-based drug design that attempts to make compounds with optimal molecular interactions.

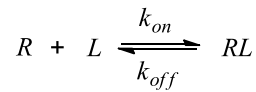
Molecular recognition processes involve agonists or antagonists with receptors, substrates or inhibitors with enzymes, antigens to antibodies. In 1894, Emil Fischer suggested the “lock-and-key” theory to demonstrate the ligand-protein interaction mechanism. Later, it was observed that active site of the protein is reshaped by interactions with the ligands and *vice versa*, so both the protein active site and the ligand itself rearranged comparing to original state during molecular recognition. The “lock-and-key” theory is no longer sufficient to know how the key fits into the lock. Accordingly, in 1958, Daniel Koshland proposed the “induced-fit model”, which is more appropriate (Jorgensen, 1991) to portray

how ligand and protein may adjust their conformation(s) to achieve an overall “best-fit” (Wei et al., 2004). Today, the importance of water molecules is starting to be recognized and addressed. Two major roles are envisioned to water molecules in ligand binding. The first role is to stabilize a protein-ligand complex through creating a hydrogen bonding network. The second one is the ability of water molecules to be displaced upon ligand binding, and there are costs to desolvate both protein and ligand.

2.1.1 Ligand-receptor interactions

2.1.1.1 Thermodynamics of ligand binding

Binding processes can be described using a simple thermodynamic equation. Reversible binding processes of a ligand (L) binding to a protein (receptor, R) is in equilibrium when concentrations remain unchanged.



The process is dynamic and depends on the association rate (k_{on}) and dissociation rate (k_{off}).

The equilibrium dissociation constant K_d is then defined as: $K_d = \frac{k_{on}}{k_{off}}$

Sometimes the binding is assessed in relation to displacement of a competitor, in which case inhibition constant K_i is measured.

Spontaneous processes takes place only when they are related with negative Gibbs free energy. For reversible binding processes at equilibrium, the binding affinity or the molar Gibbs free energy ΔG is related to the dissociation constant K_d by:

$$\Delta G = -RT \ln \left(\frac{K_d}{C^\theta} \right) = \Delta H - T\Delta S$$

in which R is the ideal gas constant, T the absolute temperature and the standard reference concentration $C^\theta = 1 \text{ mol/L}$, ΔH is the enthalpic change and $T\Delta S$ is the entropic term.

Depending on the sign and relative weight of the ΔH and ΔS components, one specific

binding process (i.e. a negative ΔG) can be either enthalpy- or entropy-driven.

- ΔH can be roughly approximated as the difference between the energy used to break bonds in a chemical reaction and the energy gained by the formation of new chemical bonds in the reaction. It includes intramolecular and intermolecular weak interactions (inside protein; protein to ligand; protein to solvent; solvent to solvent; etc.), as well as the costs of solvating and desolvating the protein and the ligand.
- ΔS , the entropy change, qualifies the molecular energy dispersed in a process. ΔS represents roughly the loss of flexibility of the ligand; the loss of flexibility of the protein; and the rearrangements in the solvent network; etc.

Experimentally, only ΔG and ΔH can be measured and can be used to determine ΔS . ΔG is measured from K_d with binding (or functional) assays; ΔH is characterized using isothermal titration calorimetry (ITC). Such data is for example available together with structural protein-ligand data about the complexes in a freely available database, SCORPIO (Olsson et al., 2008). To date, the new version SCORPIO 2.0 holds 30 different proteins, 173 ligands and 118 unique protein-ligand complexes. The aim of this database is to relate structure to thermodynamics.

In computational drug discovery, a phenomenon has been observed based on comparing the $T\Delta S$ to ΔH . Firstly, binding affinity (ΔG) is not correlated to either enthalpy or entropy separately (Reynolds and Holloway, 2011). Furthermore, it is empirically observed that when considering analogous series of compounds, increase of ΔH are related to decrease of $T\Delta S$. As a consequence, ligands with the most favorable entropies of binding actually have very unfavorable enthalpies of binding, while reciprocally the most favorable enthalpies have negative entropies of binding. This is very important point because the majority of structure-based drug design software use interaction energies for ligand pose assembly and ranking. At this stage, these energies are most closely comparable to enthalpies only. Few docking programs include entropy items and none is accurately computed (Hanzi et al., 2014).

The estimation of ΔG is difficult especially since ΔH and $T\Delta S$ are large compared to ΔG .

Large uncertainties in enthalpy measurements can yield apparent (but false) entropy-enthalpy compensation (Chodera and Mobley, 2013). In case the entropy-enthalpy compensation does occur, the main physical reasons are solvent reorganization and receptor flexibility, while they seldom are relevant to the conformational restriction of free energy decomposition.

2.1.1.2 Solvent effects in protein-ligand recognition

The hydrophobicity plays key roles in molecular recognition. Hydrophobicity (from the Greek word hydro, water; and phobic, fearing), describes the repulsion leading to exclusion between water and nonpolar substances. The hydrophobic effect involves many factors with different enthalpic and entropic contributions and it is not yet fully understood (Biela et al., 2013). Solutes including apolar solutes are surrounded by ordered water molecules, usually as a “cage”. At the molecular level, it is entropically favorable that the contact areas of these apolar solutes with ordered water are minimized and the amount of free water maximized, leading to the phenomenon of hydrophobicity. The hydrophobic effect explains why soluble proteins fold with a hydrophobic core and hydrophilic outer surface (Dill et al., 2008), and why the core of a protein is reachable for hydrophobic ligands (Whitesides and Krishnamurthy, 2005). Since receptors and ligands predominately consist of carbon and hydrogen atoms, the importance of the hydrophobic effect is considerable in biomolecular recognition.

The solvent is referred to as bulk water. Bulk water refers to water molecules that are surrounding the protein or often loosely trapped in the protein binding site. However, water molecules, and ions, can also affect directly molecular recognition. Water molecules can mediate or bridge the contact between ligand and receptor. When these water molecules are observed in X-ray crystallography they are named structural water molecules. Structural water molecules affect the binding through reshaping the protein surface or by mediating the binding with hydrogen bonds. In general, ligands that use tightly bound water bridges are smaller, less lipophilic, and less planar and have higher ligand efficiency indices (García-Sosa, 2013). Water molecules have also been found to be important in protein-protein interactions (Rajagopala et al., 2014),(Li et al., 2012).

The process of ligand recognition involves desolvation and solvation events, both at the ligand and receptor site. Some molecules are favorable to displace, some are not. Bulk water molecules are readily replaced since the release of ordered water is entropy-favorable to ligand-receptor binding. Buried water molecules can however be sometimes be favorable to displace, most often when they are forming few hydrogen bonds (Bissantz et al., 2010); but tetracoordinated waters are however difficult to displace. Conserved water molecules are replaceable, but this replacement requires very high surrogates of the interaction in terms of geometry because the enthalpy loss might not be paid off by entropy. An example of favorable displacement of water molecules is provided by HIV-1 protease inhibitors, cyclic ureas, where a cyclic ligand obtains an increase in hydrophobic interactions as well as beneficial rigidity for binding compared to the noncyclic, hydrated ligand complexes (Lam et al., 1994). Another example is carbonic anhydrase. Among a series of structurally homologous heterocyclic aromatic sulfonamides binding to carbonic anhydrase, changes in the number and organization of water molecules localized in the active site, rather than the release of structured water from the apposed hydrophobic surfaces, determine the thermodynamics of binding of these ligands to the protein (Snyder et al., 2011). In this example, gain in binding affinity is therefore enthalpy-driven.

Computational prediction of the location and thermodynamic properties of water molecules has been recognized of key importance in drug discovery (Yang, et al., 2014). The goal is to identify water molecules that are favorable or unfavorable to displace, and use this information to grow ligand parts. Computational methods based on long molecular dynamic simulations of water molecules with a restrained protein are for example used in WaterMap (Schrödinger Ltd) (Abel et al., 2008), a commercial suite. It is also able to differentially displace and retain specific water molecules in protein binding sites (Young et al., 2007). A free software based on molecular dynamics, WATsite, has been developed (Hu and Lill, 2014). Molecular docking protocols have also been used and they claim good accuracies (70-80%) in recreating the position of structural water molecules (Ross, et al., 2012).

Other protocols have for example used a geometrical scoring function named WaterScore (García-Sosa, et al., 2003), together with a ligand-superposition-based pharmacophore to distinguish bound and displaceable water in the protein binding sites of thymidine kinase

and poly-(ADP-ribose) polymerase (Lloyd et al., 2004). Other protocols have used the HINT free energy scoring function (Eugene Kellogg and Abraham, 2000) and geometric descriptor Rank (Kellogg and Chen, 2004) to predict crystallographic water molecules with accuracies of 86% (training) and 92% (test) (Amadasi et al., 2008). JAWS (Michel, et al., 2009) is a Monte Carlo simulation and statistical thermodynamics methodology to determine the position of water molecules in the binding site of a protein or protein-ligand complex.

2.1.2 Specific intermolecular interactions (noncovalent interactions)

When a ligand interacts with a receptor, a number of forces are responsible for the interaction. Though a very weak interaction individually compared to a covalent bond, cumulatively noncovalent bonds are fundamental to ligand protein interaction.

Interactions originate from the permanent or partial charges carried by atoms. When studied using classical mechanics, all electronic effects are combined and each atom is assimilated to a single particle and each particle assigned a radius and a net charge. The net charge represents its formal charge, the presence of an electron lone pairs, and the delocalization of electrons across bonds. Formal charges are the result of loss or gain of a proton H^+ (i.e. a hydrogen atom that has “lost” an electron, for example in the dissociation $-COOH \rightleftharpoons COO^- + H^+$, the $-COO^-$ group bears an “additional” electron as a result of the dissociation). Partial charges are due to excess of electrons (lone pairs) or to electronic delocalization. There are two types of electronic delocalization across bonds: inductive effects, when two bonded atoms differ in electronegativity; and resonance effects, the delocalization of electrons across conjugated systems. Dipoles, that may be permanent or temporary (induced), are created by the presence of several charged centers in a molecule. For example, the peptide bond has a permanent dipole that is oriented from the amide nitrogen to the carbonyl and 1/3 of a permanent full charge.

The principal types of interactions are summarized in Table 1. Columbic interactions are attractive or repulsive interactions formed by two charged groups and follow Coulomb's law. Dipole-dipole interactions result when the partially negative portion of one of a dipole approaches the partially positive portion of the second dipole while they are aligned in

space. Hydrogen bonds (H-bonds) are a specific type of dipole-dipole attraction that involves the interaction between a partially positively charged hydrogen atoms to a highly electronegative atom such as a partially negatively charged oxygen, nitrogen, sulfur, or fluorine atom. The Van der Waals interaction includes the Keesom force (attractive or repulsive forces between molecules other than those due to covalent bonds), Debye force (attractive interaction between a permanent multipole on one molecule with an induced multipole on another) and London dispersion force (quantum-induced instantaneous polarization or instantaneous dipole-induced dipole force). The order of bond strength is as following: ionic > H-bond > dipole-dipole > van der Waals.

A potential function or force field in molecular modeling, calculates the molecular system's potential energy (E) in a given conformation as a sum of individual energy terms.

$$E = E_{covalent} + E_{noncovalent}$$

where $E_{covalent}$ consists of three main terms: $E_{covalent} = E_{bond} + E_{angle} + E_{dihedral}$ and $E_{noncovalent}$ consists of two terms: $E_{noncovalent} = E_{electrostatic} + E_{van\ der\ Waals}$

In CHARMM force field $E_{van\ der\ Waals}$ or E_{LJ} is simplified as:

$$E_{LJ} = \sum_{nonb.pairs} \epsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^6 \right]$$

E_{elec} is given as:

$$E_{elec} = \sum_{nonb.pairs} \frac{q_i q_j}{\epsilon r_{ij}}$$

Table 1 *Molecular interactions*

Chemical bond		Strength	Formula to model the force of interaction (force)	Description
Covalent bond		100 kcal/mol	$F_{bond} = \sum_{bonds} K_b(b - b^0)^2$ $F_{angle} = \sum_{angle} K_\theta(\theta - \theta^0)^2$ $F_{dihedrals} = \sum_{dihedrals} K_\phi(1 + \cos(n\phi - \sigma))$	Geometry (length, angles) depends on atom types and hybridization
Electrostatic interaction (Coulombic)		5-40 kcal/mol	$F = \frac{1}{4\pi\epsilon^0} \frac{ q_1q_2 }{r^2}$	q_1, q_2 : signed magnitudes of the charges, r : distance between the charges non-directional
van der Waals	Keesom force	0.5-2 kcal/mol	$F = \frac{-2m_1^2m_2^2}{48\pi^2\epsilon_0^2\epsilon_r^2k_bTr^6}$	m : charge per length, ϵ_0 : permittivity of free space, ϵ_r : dielectric constant of surrounding material, k_b : Boltzmann constant and r : distance between molecules
	Debye force		$F = \frac{-m_1^2\alpha_2}{16\pi^2\epsilon_0^2\epsilon_r^2r^6}$	α : polarizability
	London dispersion force	<1 kcal/mol	$F = -\frac{3\alpha_2I}{4r^6}$	I : the first ionization energy of each molecule

2.1.2.1 Hydrogen bonds

Hydrogen bonds are the most important specific interactions in biological recognition processes. When designing compounds, hydrogen bonding should be spatially well positioned otherwise they become detrimental to activity. A hydrogen bond is an interaction between a proton donor (D) (electronegative atom, especially nitrogen or oxygen) and a proton acceptor or lone pair carrier (A). An aromatic ring can serve as a hydrogen acceptor because an electron-rich π system above and below the benzene ring hosts a partial negative charge (Du et al., 2013). Hydrogen bonds have energies in the 1-12

kcal mol⁻¹.

Hybridization of the D and A atoms and especially their ionic character ΔpK_a are the determinants of hydrogen bonds strength: When ΔpK_a approaches to zero, the strength of hydrogen bond gets close to maximum (Gilli et al., 2009). The strength of hydrogen bonds is reflected in the geometries (directionality) and interatomic distances among the donor, acceptor and hydrogen atoms. Salt bridges where the acceptor is acidic and the donor a base are an example of strong charged reinforced hydrogen bonds and are very constrained both in terms of distance and geometry.

In terms of distance, the sum of van der Waals radius between two electronegative atoms defines the length of hydrogen bond, but some have been observed to be shorter. Typical length is for example, the distance between amide C=O and OH which has a median about 2.75 Å in Cambridge Structural Database (CSD), with amide NH donors, the median distance increases to about 2.9 Å (Bissantz et al., 2010).

In terms of geometry, hydrogen bonds are generally formed along the direction of free electron pairs of the acceptor atoms. In theory, the preferred angle C=O...H is 120°, which corresponds to the lone pair direction of the carbonyl. Of important note, hydrogen bond can be bifurcated.

2.1.2.2 Halogen bonds

Halogen bonds are comparable to hydrogen bonds, but the halogen atom is shared between a donor D and an acceptor A. Halogen bonds were discovered in 1986: in crystallographic environment, the electrophiles, such as metal ions, tend to approach halogens of C-X (X = Cl, Br, I) at an angle of ~100° (“side-on”), while nucleophiles, such as oxygen and nitrogen, at ~165° (“head-on”) (Ramasubbu, et al., 1986). Halogen bonds involve a negative site, *e.g.* a lone pair of a Lewis base and the positive electrostatic potential on the outermost portions of covalently-bonded halogen atoms, termed as σ -hole (Politzer et al., 2007),(Wolters et al., 2014). σ -hole is due to the atom's charge distribution anisotropy (Politzer, et al., 2013).

Halogen bonds are important to medicinal chemistry since many drugs are halogenated molecules. An example of halogen bonds is given by phosphodiesterase type 5 (PDE5) inhibitors, shown in Figure 1. The halogens interact with the residue Tyr612 and one water molecule involved in a network. An increased IC_{50} ranged from 90.0 nM with fluorine (**2**, Fig.1), 35.9 nM with chlorine (**3**, Fig.1), 13.3 nM with bromine (**4**, Fig.1) to 7.2 nM with iodine (**5**, Fig.1) of inhibitor, respectively while reference (**1**, Fig.1) being 51.8 nM when it is hydrogen atom (Ren et al., 2014).

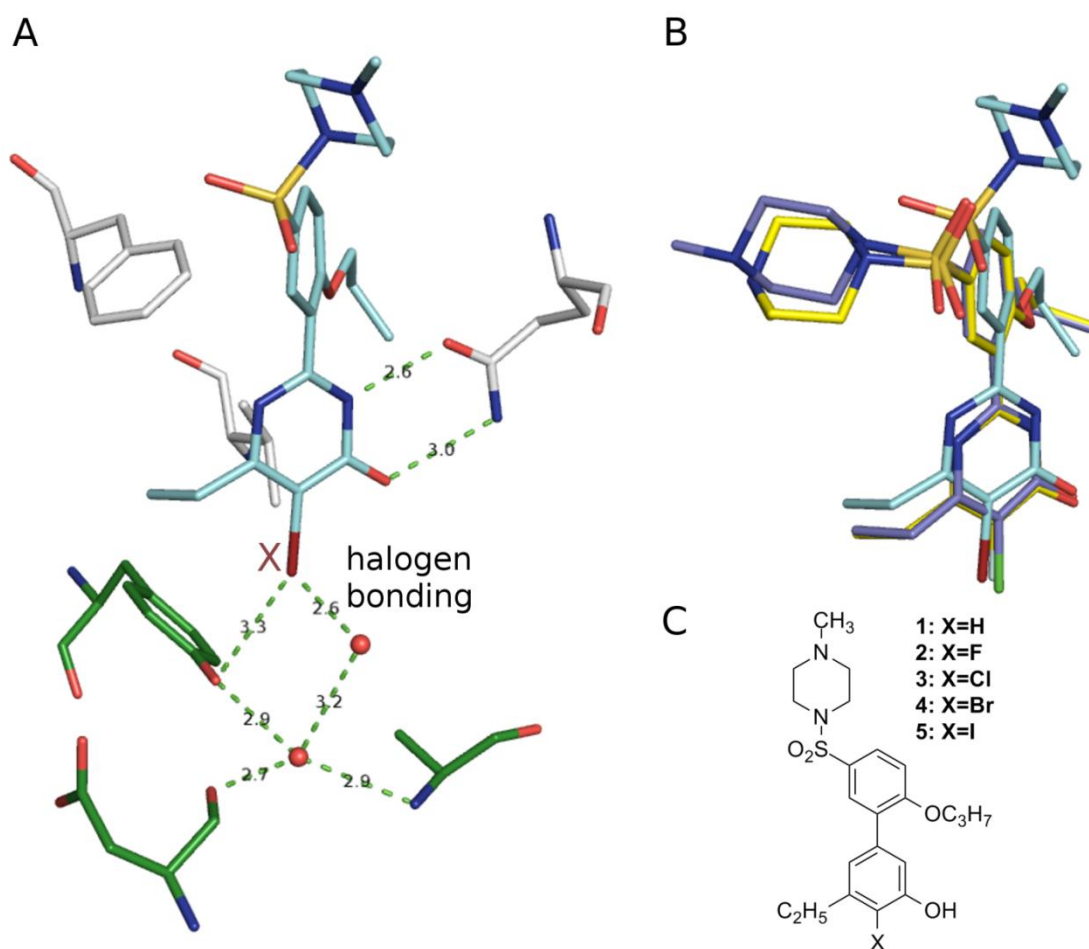


Figure 1 Structures of five inhibitors and their complexes with the catalytic domain of PDE5. (A) The detailed interactions between **4** and PDE5. (B) The superimposition of **2** (yellow), **3**(slate), **4** (cyan) by fitting the coordinates of proteins of three complex structures PDB-code: 3SHY, 3SHZ, 3SIE individually. (C) Chemical structures of the five inhibitors. Adapted from Ren et al., 2014.

2.1.2.3 π interactions

Interactions involving π systems are critical to protein-ligand recognition since the side chains of aryl-containing amino acid such as Trp, Phe, Tyr, and His are often exposed to binding site. In addition, aryl moiety is often present in small molecules. Several π interactions include π - π stacking, alkyl-aryl, and cation/anion- π . The electrostatic theory in polarized π systems can be used to explain these interactions (Martinez and Iverson, 2012). The prototype of π - π system is the benzene dimer. In the π system of benzene, a partial negative π electron density occurs above and below the plane of the ring and a partial positive charge on the periphery of the rings.

There are three types of π - π stacking interactions (Fig. 2). The parallel displaced and T-shaped modes are the most stable conformations. This charge distribution leads to favorable interactions in the T-shaped and parallel displaced conformers, but to less favored interactions in the stacked one. (Waters, 2002) However, it becomes favorable if the polarity of one ring is reversed by the addition of electron withdrawing substituents. In protein structures, the parallel-displaced arrangement is 0.5-0.75 kcal/mol more stable than T-shaped (McGaughey, 1998).

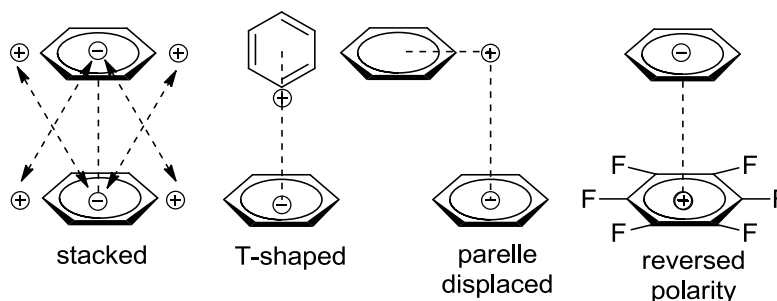


Figure 2 Geometries, quadrupole moments of typical π - π aromatic interactions. Adapted from Matthews et al., 2014.

Cation- π interaction is non-covalent intermolecular interaction between the face of electron-rich π system and cation. Energetically, a cation- π interaction is comparable to or stronger than a typical hydrogen bond. Theoretical studies also indicated typical cation- π interaction energies of about 5.5 kcal/mol in an exposed aqueous media (Gallivan and Dougherty, 2000). In the case of ligand-receptor interaction, the π system usually comes from the residue of aryl containing amino acids, while cations are supplied by the protonated amines (RNH_3^+), quaternary ammonium ions, sulfonium ions and carbocations (Dougherty, 2007). Many examples of binding sites rich in aromatic amino acids and

where cation- π interactions occur can be found, such as the GABA receptor (Lummi et al., 2005) or the D2 dopamine receptor (Torricelli et al., 2009), (Zacharias, 2002), (Dougherty, 2013).

Anion- π interaction is an interaction, where a negatively charged group with the positive electrostatic potential on the ring edge of an aromatic group forms a favorable anion-quadrupole interaction (Jackson et al., 2007), (Schottel, et al., 2008). Philip and coworkers studied anion- π pairs and found them both in α helices and β strands in Protein Data Bank (Philip et al., 2011). An example of molecular recognition is given by Schwans et al. who demonstrated that anion-aromatic interactions between the anionic Asp38 and Phe54 and Phe116 were used to position the general base in the ketosteroid isomerase active site (Schwans et al., 2013).

2.2 Chemical replacement strategies

2.2.1 The Structure-Property Principle

The Structure-Property Principle is the cornerstone of medicinal chemistry. It states that similar molecules have similar activities. It is used by medicinal chemists to suggest similar candidate compounds with certain activity; to assess how similar a candidate is comparing to commercially available one in order to avoid the intellectual property conflict; and to investigate whether structurally dissimilar candidates are also reliable leads and what is the mechanism behind these novel structures (Maggiore et al., 2013). The marketed drugs are mimics of endogenous molecules to varying degrees. This may spur one to learn from these natural products because most of them were readily recognized by macromolecules with decent binding affinity, such ADP analogs.

2.2.1.1 Defining molecular similarity

Depending on individual perspective, “similarity” has many different meanings et al., 2004), (Willett, 2013). One aspect of “similarity” is based on physicochemical characteristics, such as log P , boiling point, molecular weight, solubility, electron density

and dipole moment. Another point of view is to describe molecular similarity based on features such as molecular fingerprints, ring system and topology, as is done in chemoinformatics. A third viewpoint is to consider similar compounds with approximately the same atomic constituents, functional groups, backbone, or ring systems. Not very often, compounds that can be obtained by a single step of synthesis may be considered similar. Böhm et al. thought two scaffolds can be different if they were synthesized using different synthetic routes, no matter how small the change might be (Böhm, et al., 2004). Biological similarity is another concept that disregards the structural features of the compounds; instead, it relates the activity of compounds against a panel of biological targets.

2.2.1.2 Activity cliffs

Structure-activity relationship (SAR) discontinuity, i.e., defined as pairs of structurally similar compounds with a significant difference in potency, has been referred to as activity cliffs (Maggiore, 2006). On average, one in five compounds was shown to be involved in activity cliff (Sisay, et al., 2009). This makes SAR prediction difficult, but on the other hand, activity cliffs found in compound data sets might be a potential starting point of novel compounds (Dimova et al., 2013). One difficulty in predicting the effect of modifications arises since binding modes of structurally similar ligands may vary. Based on an investigation of the PDB, Boström et al. attribute this uncertainty to reorganization of water molecules and side-chain but seldom backbone movements of protein (Boström, et al., 2006).

2.2.2 Isosteres and bioisosteres

Bioisosteric replacements are one of the important strategies in the design of new molecules in medicinal chemistry, often utilized in lead optimization process with the aim of mimicking or improving properties such as pharmacokinetics, metabolism, solubility, or reducing side effects while keeping or improving potency.

2.2.2.1 The development of bioisosterism

About one century ago, the term isosterism was coined by Irving Langmuir (Langmuir,

1919). His work focused on the relationships between electronic configuration and physical properties. For example, carbon monoxide and the cyanide ion have the same number of electrons in their valence shell and are similar in term of physical properties, such as critical pressure, critical temperature, viscosity, magnetic susceptibility, but not freezing point. Later, H. G. Grimm introduced the concept of *pseudoatom*, which hypothesizes similar properties for atoms with an added hydrogen to atoms with the next atomic number. Hence, OH is isosteric with F.

Following the work of Erlenmeyer, Thornber proposed a general definition of bioisosteres: “*Bioisosteres are groups or molecules which have chemical and physical similarities producing broadly similar biological properties*” (Thornber, 1979). He thought eight aspects can be considered when one tries to identify structural alterations: size, shape, electronic distribution, lipid solubility, water solubility, pK_a , chemical reactivity and hydrogen bonding capacity. It was later stated that bioisosteres are not necessarily alike in their physical but biological properties (Burger, 1991).

2.2.2.2 Phosphate isosteres

At biological pH (pH 7.4), phosphate is predominately charged, with pK_{a1} 1.54 and pK_{a2} 6.31. Phosphate shows tetrahedral molecular geometry. Instinctively, bioisosteres of phosphate should be fragments that are similar to phosphate in terms of charge, hydrogen bonding and geometry. Typical phosphate bioisosteric replacements have been reviewed by Elliott et al. (Elliott et al., 2012) and Rye et al. (Rye and Baell, 2005). These are phosphonates, carboxylates and malonates (Desvergnès et al., 2012), sulfamates (Moreau et al., 2013), squaramides and squaric acids (Niewiadomski et al., 2010), boron-containing fragments (Albers et al., 2011), and phosphorothioates (Liu et al., 2008). Many studies have explored different types of replacements for phosphate. Some of these examples are discussed below.

Phosphorus-containing isosteres

Structurally, the simplest way to obtain a phosphate bioisostere is to replace the bridging oxygen of a phosphate by methylene, leading to the phosphonate moiety. An example of a

phosphonate replacement is given by the study of He et al. where two series of metabolically stable phosphonate bearing analogs were synthesized (He et al., 2011). All analogs (Fig. 3) bind the same four-phosphate-adaptor protein 1 pleckstrin homology domain, where natural compound PtdIns(4)P do with improved K_d from 0.46 μM (compound **6**, Fig.3) to 0.39 μM (compounds **9** and **10**, Fig.3) and 0.25 μM (**7**, **8**, Fig.3).

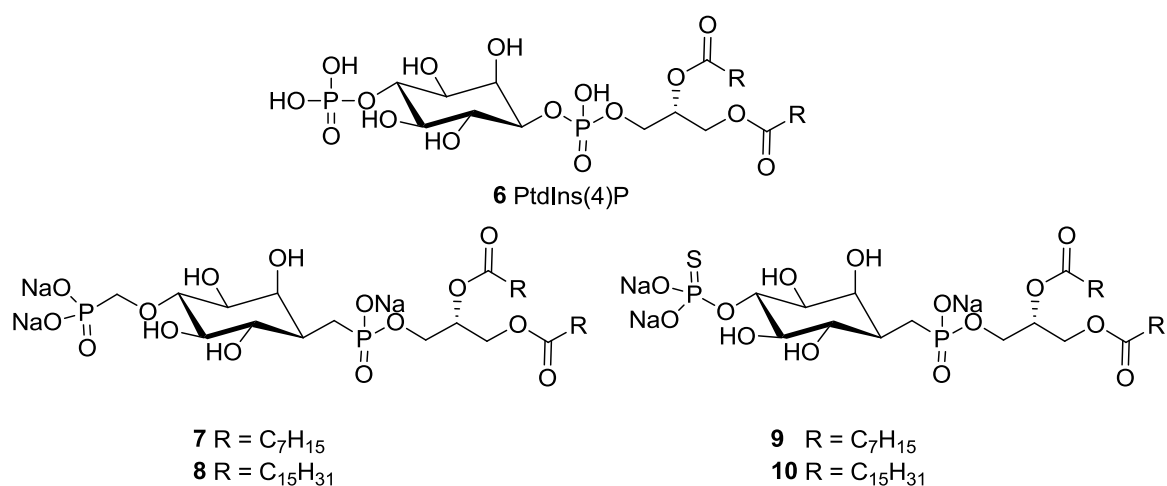
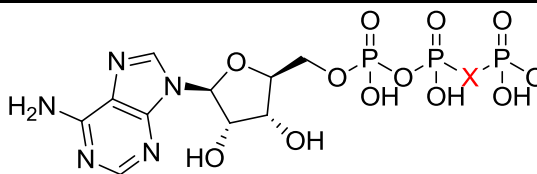


Figure 3 The metabolically stable analogs PtdIns(4)P-MP (**7-8**) and PtdIns(4)P-PT (**9-10**) of PtdIns(4)P (**6**). Adapted from J. He et al., 2011.

The phosphorus-carbon bond in phosphonates is more stable to hydrolysis than phosphorus-oxygen in diphosphate. Phosphonate compounds have pK_a that varies between 6.7 and 8.4. (Table 2) Though the stability of phosphates is improved by phosphonate, phosphonate is not a perfect bioisosteric replacement of phosphates in terms of pK_a . The second pK_a of phosphonate (7.49) is significantly higher than for phosphate (6.31) indicating that phosphate will be dianionic in biological pH, while phosphonate will be mono-ionized. This has driven efforts to develop phosphonate analogs with lower pK_a such as α -fluorophosphonate or bromophosphonate (Elliott et al., 2012), (Tulsi, et al., 2010). The C-Br bond length (1.93 Å) is longer than C-F (1.41 Å), and an α -bromophosphonate group ($pK_{a2} \sim 6.52$) is 10-fold more basic than an α -fluorophosphonate moiety ($pK_{a2} \sim 5.6$), which may offer α -bromophosphonate containing ligands some advantage over α -fluorophosphonate ones. The synthesis and biological activity of α -bromophosphonates has been reviewed and discussed by Downey and Cairo recently (Downey and Cairo, 2014).

Table 2. ATP analogs **11-16** bearing terminal phosphate isosteres and their respective pK_a values. Adapted from Elliott et al., 2012



Compd	X	pK _a
11	O	7.1
12	CH ₂	8.4
13	NH	7.7
14	CHF	7.4
15	CF ₂	6.7
16	CCl ₂	7.0

Carboxylic acid

One of the most commonly used non-phosphorus isostere of phosphate is the carboxyl group (Ballatore, et al., 2013), (Sirivolu et al., 2012). Furthermore, carboxyl groups are widely represented in drug molecules with about 450 carboxyl-containing drugs marketed worldwide (Ballatore et al., 2013). Nevertheless, the presence of carboxyl groups also leads to potential drawbacks, including low metabolic stability, toxicity, and a difficult passive diffusion across cell membranes, particularly in the context of central nervous system (CNS) drug discovery (Pajouhesh and Lenz, 2005).

Carboxylic acids have pK_a that ranges from 0.5 to 4.2 depending on substituents (Jinhua, et al., 2006) and in the ionized form they contain two hydrogen bond accepting groups. For example, changing a phosphate to a carboxyl in the design of sphingosine-1-phosphate (S1P) receptor antagonist led to an active compound **18** with 19 nM activity, to be compared to ~1 nM for the parent endogenous ligand **17** (Fig. 4) (Högenauer, et al., 2010).

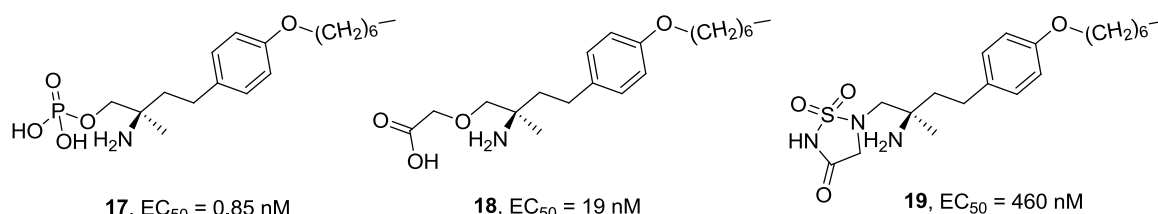


Figure 4

S1P receptor antagonists. Adapted from Högenauer et al. 2010.

Sulfur-containing isosteres

In terms of pK_a and tetrahedral shape, phosphate can be bioisosterically replaced by sulfate and sulfonate. Sulfate and sulfonate have pK_a in the range -3.4 to -1.9 and are thus fully ionized at physiological pH. At the S1P receptor, changing phosphate to sulfonylhydantoin-containing compound **19** (Fig. 4) led to a compound with an activity of 460 nM (parent compound **17**, ~1 nM). Another example is reported by Hussain et al. for the protein-tyrosine phosphatase 1B. While the parent inhibitor has an IC_{50} of 8 nM (compound **20**, Fig.5, containing difluoromethylenephosphonic acid), the replacement of difluoromethylenephosphonic acid with difluoromethylenesulfonic acid gives rise to compound **21**, which is about 1000-fold less potent ($IC_{50} = 13 \mu M$) (Hussain et al., 2008).

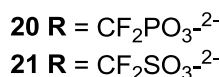
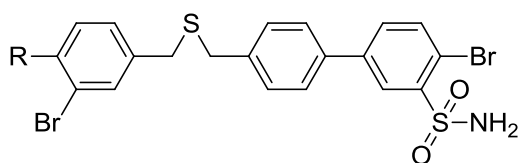


Figure 5 Representative protein tyrosine phosphatase 1B inhibitor structures. Adapted from Hussain, Ahmed et al., 2008.

Other sulfur-containing replacements are methyl sulfamate, for which pK_a is 15.9 and acylsulfonamide with a calculated pK_a of 2.5 (He, et al., 2006). Sulfamate therefore is likely to be a neutral phosphate isostere.

Boronic acids isosteres

Boron containing compounds are attractive for drug development (Bross et al., 2004), (Shapiro, 2009). Boron containing phosphate isosteres are boronic acids, benzoxaboroles and boranophosphates (Fig. 6). The pK_a of boronic acids is 8.8, while that of benzoxaboroles is between 7.34 and 8.32 depending on the structure of the substituent in the 3-position (Table 2). Comparing to simple aryl boronic acids, such as phenylboronic acids, benzoxaboroles are more hydrolytically stable (Snyder, et al., 1958), Boranophosphate is more hydrophobic but still water soluble (for a review see (Elliott et al., 2012)).

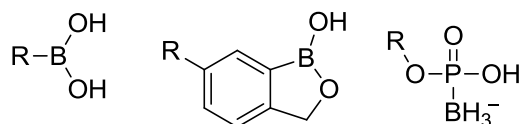
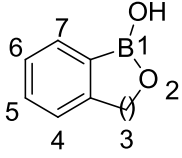


Figure 6 Structures of boronic acids, benzoxaboroles and boranophosphates

Table 2 pK_a of Benzoxaboroles Adapted from Tomsho et al., 2011

		
compd	3-position	pK_a
benzoxaborole	$-CH_2-$	7.34 ± 0.02
benzoxaborin	$-CH_2-CH_2-$	8.40 ± 0.03
3,3- <i>gem</i> -dimethyl	$-C(CH_3)_2-$	8.32 ± 0.03

Squaramide-based isosteres

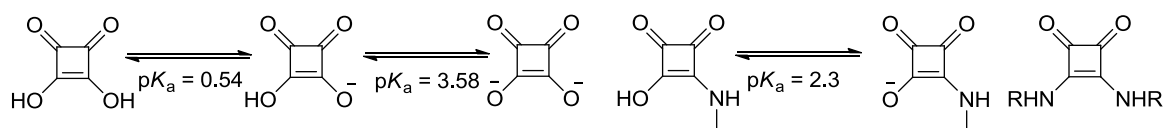
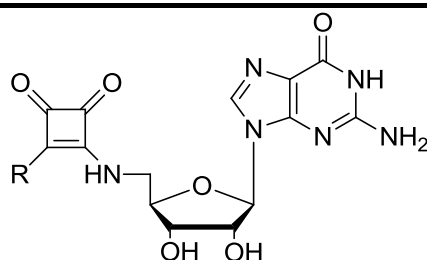


Figure 7 The pK_a value of squaric acid and squaramide. Adapted from Elliott et al., 2012.

The squaramide moiety has also been studied as a phosphate isostere. Squaramide is prepared from squaric acid, which has pK_{a1} of 0.54 and pK_{a2} of 3.58 (see **Fig. 7**), suitable feature for phosphate bioisostery. Like phosphate, the two secondary amines of squaramide can indeed coordinate divalent metal ions, such as Mg^{2+} . An example is found with the mannosyltransferase dolichol-phosphate mannose synthase; it is an Mg^{2+} -dependent enzyme with a GDP-mannose binding site. In a study by Niewiadomski et al. (Niewiadomski et al., 2010), the squaramide core was used to replace the diphosphate of GDP to give rise to eight GDP analogs. Regrettably, all synthesized compounds and guanosine showed only low or moderate inhibition, even for the most potent one the 2-carboxypropylamino replacement having 64.0% of inhibition residual. (Compound **29**, Table 3)

Table 3 Biological activity of squarylamides and reference compounds. Adapted from Niewiadomski et al., 2010.



Compound	R	Residual dolichol-phosphate mannose synthase activity (%)	ED ₅₀ /mM
GDP	-	8.5±1.8	nd*
Guanosine	-	97.4±4.5	nd
22	ethoxy	87.1±0.2	1.1±0.2
23	<i>n</i> -butylamino	81.3±0.3	1.0±0.1
24	nitrobenzylamino	78.0±2.0	0.73±0.2
25	2-nitrobenzylamino	84.1±5.2	>1.5
26	3-nitrobenzylamino	69.6±2.1	>low solubility
27	4-nitrobenzylamino	65.1±3.0	>low solubility
28	carboxypropylamino	91.9±5.6	0.99±0.2
29	2-carboxypropylamino	64.0±1.0	1.0±0.2

* Not determined. ED₅₀: The dose of a drug that is pharmacologically effective for 50% of the population exposed to the drug.

Other acidic replacements of phosphate

Many isosteres of carboxyl group have also been shown to work well as phosphate isosteres because of the high similarity in the properties to phosphate. A typical example of carboxyl/phosphate isostere is 1*H*-tetrazole (Malik et al., 2014). The p*K*_a of 1*H*-tetrazole is 4.90, while acetic acid has a p*K*_a value of 4.76. Allen et al showed that the carboxyl/tetrazole and carbonate/tetrazolate pairs exhibit very similar H-bonding environment and H-bond attractive energies are also highly similar (Allen et al., 2012). The relative bigger volume of tetrazole may lead however to slightly different receptor interactions (Fig. 8). Different synthetic methods and biological importance of tetrazoles have been reviewed recently (Malik et al., 2014).

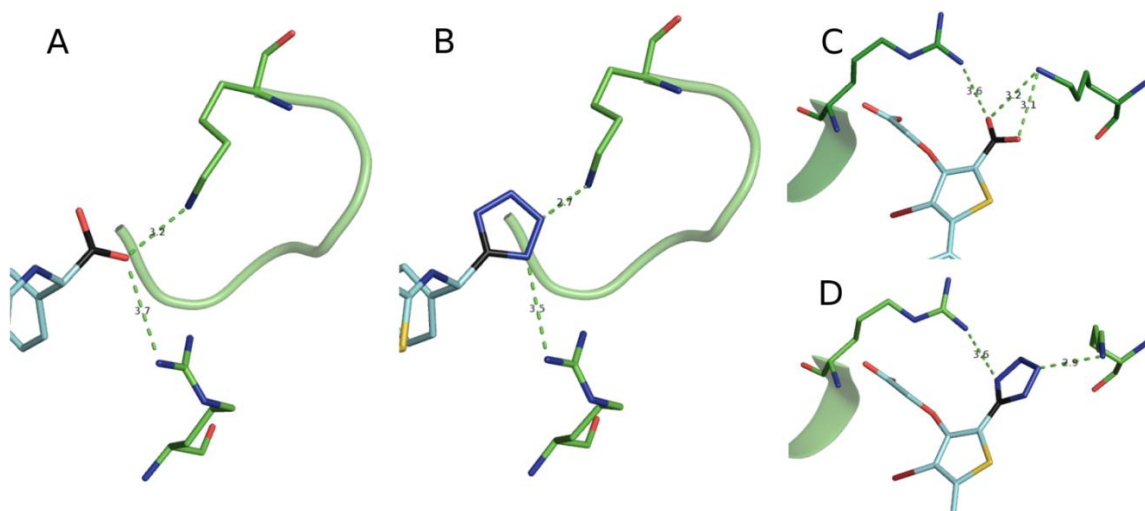
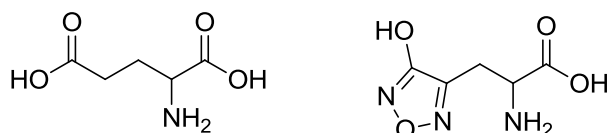


Figure 8 Superimpositions of the ligand binding sites for the tetrazolate/carboxylate ligand pairs, the two pairs of ligands are as follows (tetrazole and carboxylate ligand codes and PDB structure code(s) in parentheses) (A) VRX in 2HWI (B) VXR in 2I1R bound to hepatitis C NS5B polymerase and (C) 527 in 2QBP (D) 902 in 2NT7 bound to tyrosine phosphatase 1B. Adapted from Allen et al., 2012.

There are several less common examples of carboxyl/phosphate replacements that have been reported. For example, five-membered heterocycles such as furazan-3-ol, compound **30** (Fig.9), showed comparable K_i value to the endogenous ligand glutamate to ionotropic glutamate receptors 2, but more acidic in term of pK_{a2} (Lolli et al., 2006), (Lolli et al., 2010).



Asp, $pK_{a2} = 4.2$, $K_i = 0.28$ **30**, $pK_{a2} = 3.5$, $K_i = 0.25$ (0.043)

Figure 9 Ionotropic glutamate receptors 2 inhibitors. Adapted from Lolli et al., 2010.

There are several less common examples of carboxyl/phosphate replacements that have been reported. For example, five-membered heterocycles such as furazan-3-ol, compound **30** (Fig. 9), showed comparable K_i value to the endogenous ligand glutamate to ionotropic glutamate receptors 2, but more acidic in term of pK_{a2} (Lolli et al., 2006), (Lolli et al., 2010).

Another example is substitution by 2,6-difluorophenol that has a pK_a of 7.1. 2,6-Difluorophenol was used to develop lipophilic γ -aminobutyric acid (GABA) aminotransferase inhibitors. (Fig. 10, compounds **31** and **32**) (Qiu, et al., 1999). The

carboxyl group was also shown to be exchangeable by amide and hydroxamate groups at the GABA_A receptor (Locock et al., 2013) and KCNK potassium channels (Bagriantsev et al., 2013).

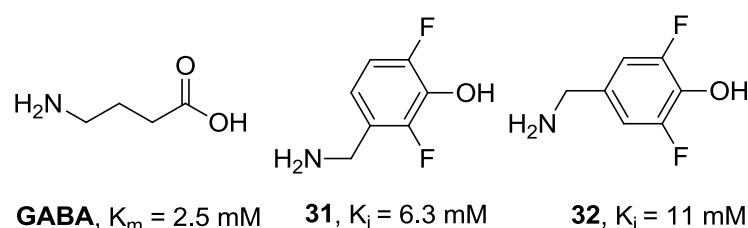


Figure 10 *γ-aminobutyric aminotransferase inhibitors. Adapted from ref Qiu et al., 1999.*

2.2.3 Scaffold hopping

Another method in lead optimization, twin of bioisosteric replacement, is scaffold hopping (Langdon, et al., 2010). Scaffold hopping assumes that the same biological activity can be achieved by maintaining some essential features in otherwise structurally different molecules. In scaffold hopping the core structure of the molecule is replaced with another scaffold that retains or improves a set of properties. The main categories of scaffold hopping are heterocycle replacements, ring opening or closure, and peptidomimetics (for review see (Sun, et al., 2012)).

2.2.3.1 Heterocyclic replacements

Heterocycles are ring structures where one or more carbon atoms of the parent carbocycles are replaced with more electronegative atoms such as oxygen, nitrogen and sulfur. For example, the biarylheterocyclic cyclooxygenases 2 (COX-2) selective inhibitors (Flower, et al., 2003), (Liedtke et al., 2012) presented in Figure 11 are different from each other mainly in the backbone heterocyclic rings, but their activity levels against COX-2 are comparable (DeWitt, 1999).

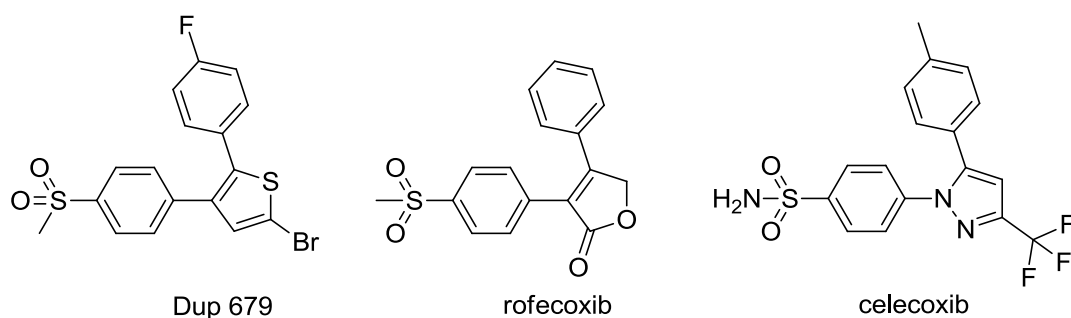


Figure 11 *Biarylheterocyclic scaffold hopping in cyclooxygenase 2 inhibitors.*

Their pharmacology is however different (see for review FitzGerald et al., 2001) (FitzGerald and Patrono, 2001). The first inhibitor Dup 679 (Gans et al., 1990) failed to reach the market. Modifications based on Dup 679 led to the inhibitor rofecoxib, which marketed, but withdrawn due to increase in the risk of cardiovascular events. Further trial led to celecoxib that is still in use to treat patients with rheumatism and osteoarthritis.

2.2.3.2 Ring closure and opening

Ring system commonly exists in drug-like molecules, so novel scaffolds can be created through ring closure strategy. Intuitively, ring closure will decrease the flexibility of the molecule, whose binding becomes entropically less favorable for ligand binding to target. It is on the other hand easier for conformationally restrained molecules to penetrate membranes (Goetz et al., 2014), (Vieth et al., 2004). An example of macrocyclization is provided by the angiotensin IV peptide (Fig. 12), giving rise to the more potent compounds **33** and **34** to insulin-regulated aminopeptidase (Andersson et al., 2011).

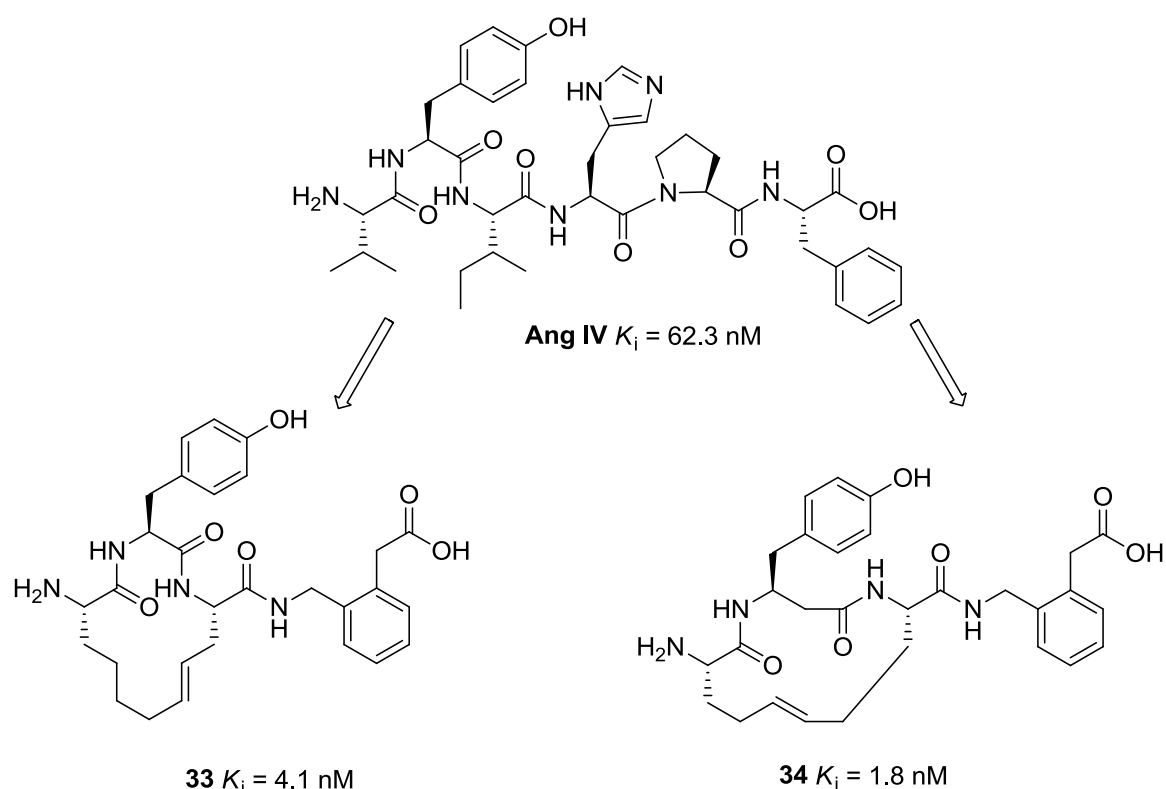


Figure 12 *Insulin-regulated aminopeptidase inhibitors. Adapted from Andersson et al. 2011.*

Among ring closure ideas, non-covalent ring structure closing through intramolecular hydrogen bonds is widely exerted and observed. When a donor and an acceptor are on the same molecule close enough, a dynamic equilibrium may exist between intramolecular hydrogen bond conformations that may facilitate the crossing of membranes, and open conformations in which the polar groups expose to solvent or approach to binding site. Intramolecular hydrogen bonding within ring systems has been studied by Kuhn et al. (Kuhn, et al., 2010) An example of selective neuronal nitric oxide synthase inhibitors bearing intramolecular hydrogen bonds is presented in Figure 13, (Labby et al., 2012) (see also PDB code: 3TYM).

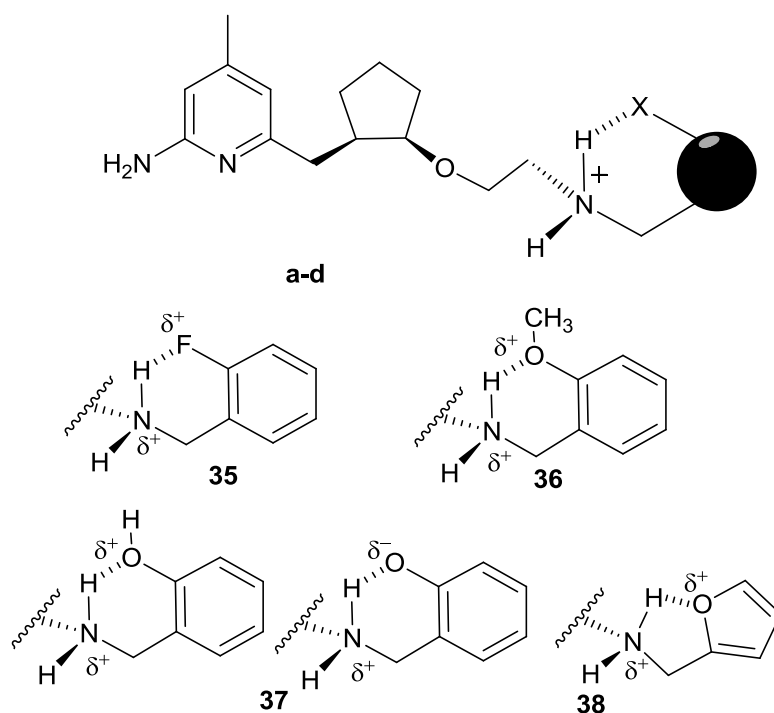


Figure 13 Chemical structures of (*S,S*)-35-38. The putative intramolecular hydrogen bond is dashed. The hydroxyl of 37 may show different conformations, in which it acts either as a hydrogen bond acceptor or a donor, sharing the H^+ in case it is hydrogen bond donor. Adapted from Labby et al., 2010.

2.2.3.3 Peptide mimetics

Structural replacement of the peptide bond has been widely studied, since peptide bonds may be readily hydrolyzed by protease once they are orally administered. Peptide-containing molecules require alternative routes of administration such as arterial injection or complicated drug delivery system, for instance porous silicon nanoparticles to transport drug to target organs (Liu et al., 2013). Alternatively, medicinal chemists prefer to use structural modifications, including the cyclization of polypeptides discussed in the previous section (Fanelli et al., 2014). Cyclic peptides have favorable properties, such as increased resistance to proteolysis, specificity and constrained conformations relative to linear ones (Menegatti et al., 2013), and they are therefore promising candidates for synthetic affinity ligands (Wels et al., 2005). Multiple novel strategies for cyclization, such as metal-ion-assisted, sulfur-mediated cyclizations, ring-contraction involving lactones, azide-alkyne cycloadditions and ring-closing metathesis have been applied (White and Yudin, 2011). Replacement of an amide group with an isosteric cyclic group has also been widely used, since compounds with multiple heterocycles are able to mimic

the orientation and disposition of key residues at protein interfaces (Lao et al., 2014). Advances in the synthesis of conformational restricted building blocks and peptide bond isosteres have been reviewed (Vagner, et al., 2008).

Small molecule mimetics of protein structures have also been designed (see a review of Ross et al. (Ross, et al., 2010). Protein-protein interactions and protein-peptide interactions are regulated by highly ordered secondary structure motifs, such as α -helices, β - and γ -turns, and β -sheets, so small molecule mimetics of proteins try to mimic these structures. Floris et al. have reviewed computational methods to design novel peptidomimetics (Floris and Moro, 2012). In addition, peptoids, β -peptides and stapled peptides are commonly used to alter the peptide properties. Peptoids are a class of peptidomimetics that have side chains linked directly to the peptide backbone nitrogen atoms rather than to the α -carbon, as they usually are in amino acids.

2.2.4 Other considerations

2.2.4.1 *Isosteric replacements are target-dependent*

An important consideration about bioisosteric replacements is that while they prove effective in one circumstance they do not guarantee the success in another setting-up (Meanwell, 2011). For example, one very successful application of carboxyl and tetrazole bioisosteric interchange is for the angiotensin II receptor antagonist losartan (**40**) (Carini et al., 1991). The tetrazole moiety in losartan offers 10-fold increase in potency compared to the carboxylic acid analog **39**. However, in a similar replacement used to develop cPLA₂ α inhibitors, the novel tetrazole-containing analog **42** is 31-fold less active than the previously reported **41** (Hess, et al., 2007).

Similarly, isosteric surrogates may lead to different pharmacological effects. The ester group containing procaine (**43**) is a local anesthetic drug, while replacement of the ester moiety with an amide group gives rise to procainamide (**44**), which is an antiarrhythmic agent. In another example, compound **46** was synthesized such that the amide bond replaces the ester bond in bevirimat (**45**), consequently leading to a significantly decreased antiviral activity of **46** against HIV-1_{IIIB} (Qian et al., 2009).

Hypothetically, structural replacements should make almost identical interactions if possible. However, given the concept of induced-fit in molecular recognition as well as the presence of other components such as water and metal ions, “bioisosteric pairs” do not necessarily take the exact same position in the binding site. The problem that specific bioisosteric replacements are prior to others for a given target family has been discussed by Wassermann and Bajorath (Wassermann and Bajorath, 2011).

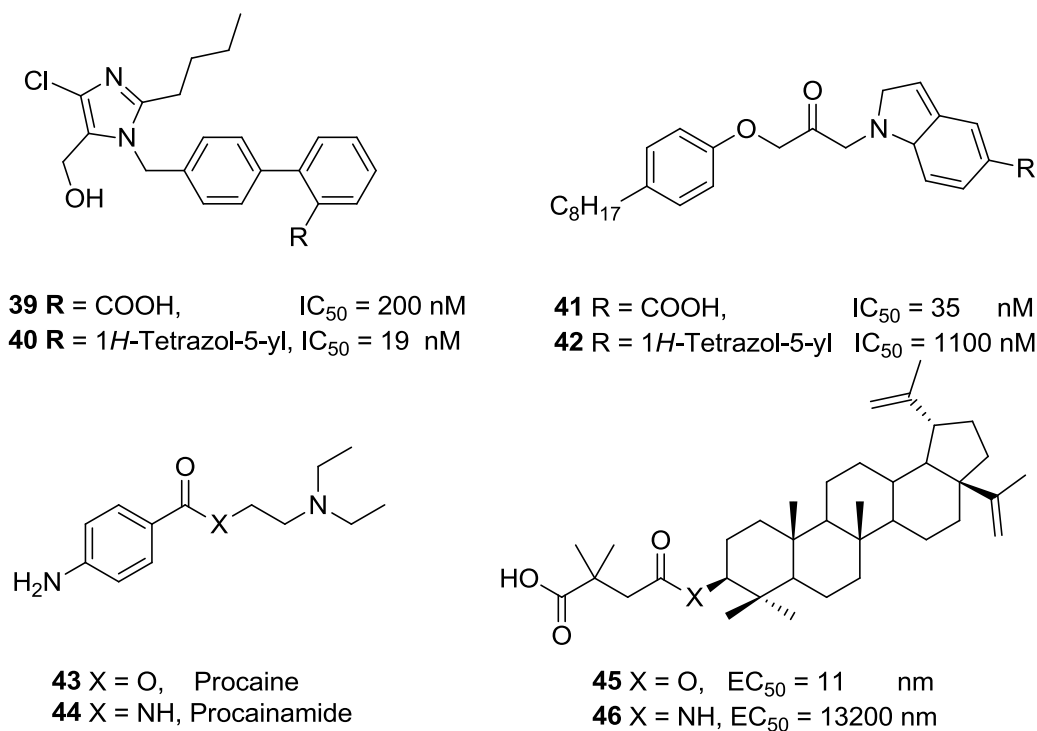


Figure 14 Structures compiled from Hess et al. 2009 and Wassermann, et al. 2011.

2.2.4.2 Prodrugs to deliver isosteres

In some cases, polarity may hamper the permeability of molecules while it is necessary for activity. The prodrug approach can address the problem (Biot et al., 2003). For example, the *in vitro* and *in vivo* result difference of naphthoquinone carboxylic acid (Fig. 15) against *Plasmodium falciparum* led to a hypothesis that the penetration due to the presence of a carboxyl group is the key (Davioud-Charvet et al., 2001). Consequently, the ester and amide-masked compounds **48** and **49** were prepared and both showed improved activity. Replacement of the carboxyl function by a tetrazolyl group increases compound bioavailability, but tetrazoles may also need to be protected to facilitate the molecules across the membranes. For example, the cyanoethyl-protected tetrazole **50** is more potent than the parent compound with a free tetrazole moiety (Biot and Bauer, 2004).

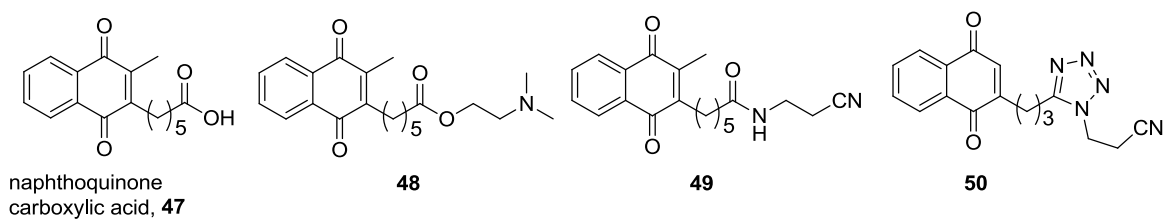


Figure 15 *Prodrugs of the naphthoquinone carboxylic acid derivative **47** protected as ester (**48**) or amide (**49**), and **50** with the 1H-tetrazolyl moiety protected with a cyanoethyl group.*

2.3 Molecular similarity in a chemoinformatics perspective

Computational modeling has gained a major importance to the drug discovery process in the last fifty years. The techniques used are divided into structure-based methods that take advantage of a three-dimensional structure to design or screen for small molecules, and ligand-based methods that aim at retrieving analogs and predict the activity of unknown compounds (QSAR modeling).

2.3.1 Databases to study isosteric replacement

Databases storing information extracted from chemical structures and their activities, or chemical structures in complex with a protein target, represent very important tools to study isosteric replacements as well as to define molecular similarity. Several databases have been developed. Bioisosteric data require at least one pair of compounds (analogs differing by one functional group) as well as activity data. An example of a commercial database is BIOSTER, which contains 26,000 pairs of potentially bioisosteric replacements (Hayward, 2012). Among the freely available databases, Swissbioisostere automatically reports potential bioisosteric replacements extracted as ChEMBL matched pairs (Wirth et al., 2013); fragments are extracted using the Hussain and Rea algorithm that cut single bonds not part of ring systems nor resonant functional groups such as amide or carboxylic acids.

The second type of databases contains statistics/information about molecular contacts but does not use biological activities. sc-PDB-Frag fragments protein-bound ligands and computes the interaction of each fragment with their protein environment, then sorts these protein-ligand interaction patterns similarly to the reference protein-ligand complex

(Desaphy and Rognan, 2014). Derived from both CSD and PDB, the commercial database IsoStar contains information of molecular interactions. This information on the frequencies and directionalities of intermolecular contacts is valuable to medicinal chemists interested in identifying bioisosteric replacements (Bruno et al., 1997).

Another class of prediction of bioisosteres is based on chemical similarity alone. pepMMsMIMIC, a web-oriented peptidomimetic virtual screening tool, suggests which chemical structures are able to mimic the protein-protein recognition of the input peptide using pharmacophore and shape similarity techniques encoded into fingerprints (Floris et al., 2011). ROCS is also a shape comparison application for virtual screening based on the idea that molecules have similar shape if their volumes overlay well (Hawkins, et al., 2007). Ring systems characterized by size, shape, pharmacophore features and ADME descriptor similarity can also be used for bioisosteric design and scaffold hopping (Ertl, 2012).

2.3.2 Ligand-based drug discovery: analog search and QSAR

2.3.2.1 Representing molecules to measure their similarity

Unlike chemical structure searches, structural similarity searches look for chemical structures that are related with the query molecule, not to exact matches. Researchers can decide the similarity degree which the hit molecules should retain. This threshold can be conveniently changed according to search and biological test results. Depending on whether the molecule is represented in 2D or 3D, 2D and 3D descriptors can be used to evaluate the molecular similarity. Given the relative degree of importance of the descriptors to similarity assessment, weighting schemes can be assigned. Most of 2D descriptors, such fingerprints based on graph information are easy to retrieve, allowing fast and efficient large scale calculations.

Computationally, chemical similarity can be measured in terms of several perspectives depending on the molecular representation used:

- 1) Topological similarity (molecule represented as a graph)
- 2) Numerical descriptor-based (vector of numbers)

- 3) Fingerprint-based (bitstring)
- 4) Pharmacophoric (small set of key features and distances)
- 5) Field-based (grid/array typically storing information about shape and electrostatic)

2.3.2.2 Adjacency/connectivity matrices and graph-based representation

Molecules are composed of connected atoms through covalent bonds. Atoms are characterized by their types (C, O, N, etc) as well as their hybridization states (C.sp2, C.sp3 etc). Bonds between atoms are characterized by different bond order, for example 1 for a single bond, 2 for a double bond and 1.5 for bonds in conjugated benzene. A simple way to represent molecules is through an adjacency or connectivity matrix, as for example in Figure 16.

	Adjacency matrix							
	1	2	3	4	5	6	7	8
	1	0	1	0	0	1	1	0
	2	1	0	1	0	0	0	0
	3	0	1	0	0	0	0	1
	4	0	1	0	0	0	0	0
	5	1	0	0	0	0	0	0
	6	1	0	0	0	0	0	0
	7	1	0	0	0	0	0	0
	8	0	0	1	0	0	0	0
	↓							
	Connectivity matrix							
	1	2	3	4	5	6	7	8
	1	6	1	0	0	1	1	0
	2	1	6	1	2	0	0	0
	3	0	1	8	0	0	0	1
	4	0	2	0	8	0	0	0
	5	1	0	0	0	1	0	0
	6	1	0	0	0	0	1	0
	7	1	0	0	0	0	0	1
	8	0	0	1	0	0	0	1
	↓							

Figure 16 Adjacency and connectivity matrices for acetic acid. Rows and columns correspond to the atom numbering given on the left. Off-diagonal elements of both matrices represent order of the bonds between atoms pointed by row and column. The diagonal of the connectivity matrix represents the atomic number. Adapted from <http://ccl.net/chemistry/resources/overview/index.shtml> (accessed 2.11.2014).

Chemoinformatics often use instead of these matrix graph representations. Typical tasks solved through graph isomorphisms are substructure search, for example using the Ullman algorithm (Ullmann, 1976) to identify the maximum common substructure, which is related to the extent of similarity in the compound topology (Stumpfe, et al., 2014). Another algorithm is the matched molecular pairs algorithm (Leach and Jones, 2006), (Cumming et al., 2013), which commonly used to find analogs differing by a single substitution and by this mean study chemical similarity (Kramer and Fuchs, 2014).

2.3.2.3 Numerical molecular descriptors

Descriptors are numerical values that describe the molecules. Usually 1D descriptors, for example the molecular mass, are properties of the structure; 2D descriptors are obtained from the 2D structure; 3D are obtained from 3D structure (some values may then be conformation-dependent). VolSurf, Dragon and JOELib are typical softwares for numerical descriptors calculation. VolSurf compresses the information present in 3D maps into few 2D descriptors for optimization of pharmacokinetic properties (Cruciani et al., 2000), Dragon 6 can calculate up to 4885 molecular descriptors and JOELib contains two main types: native value and atom property descriptors.

The application of descriptors is to visualize the chemical space for example using data transformations such as principal component analysis, multidimensional scaling or a neural network self-organizing map, as well as other clustering methods. The key idea is that molecules with similar descriptors will locate nearby.

Another application of numerical molecular descriptors is QSAR modeling: a statistical model relating a set of stimuli variables (X) to the potency of the response variable (Y). QSAR is derived by finding a relationship between chemical structures and biological activity in training data-set. Descriptors thought suitable to describe the relationship are selected and then some form of statistical analysis is performed to generate a model, which is validated with a test data set. Unlike physical methods, such as molecular dynamics, empirical QSAR methods do not rely on any understanding of the physics.

Though QSAR model is a powerful tool for drug discovery (Golbraikh et al., 2014), a lot caution should be taken in order to avoid misuses and misunderstandings of this technique (Doweyko, 2008). First off all, erroneous association of correlation with causation should be figured out. Validation is a crucial aspect of any QSAR modeling. Often, a high value of leave-one-out cross-validated R^2 ($LOO\ q^2 > 0.5$) is considered as a proof of the high predictive ability of the model. However, some studies show that a high $LOO\ q^2$ value do not necessarily indicate a predictive model and external validation is the only way to guarantee reliable QSAR models (Golbraikh and Tropsha, 2002).

2.3.2.4 Molecular fingerprints

Molecular fingerprints are a specific type of molecular descriptor that encodes molecules into bitstrings. Molecules are split into fragments and presence/absence is recorded as 1/0. UNITY 2D and fragment-based Daylight fingerprint are well-known examples. There are two types of fingerprints which the Daylight system provides. One is the “difference” fingerprint, which is used for the bond changes which occur during a reaction. Another is the “normal” structural fingerprint, which is useful as a superstructure search screen. If a query molecule is a substructure of a target molecule, then all of the bits set in the query molecule will also be set in the target molecule.

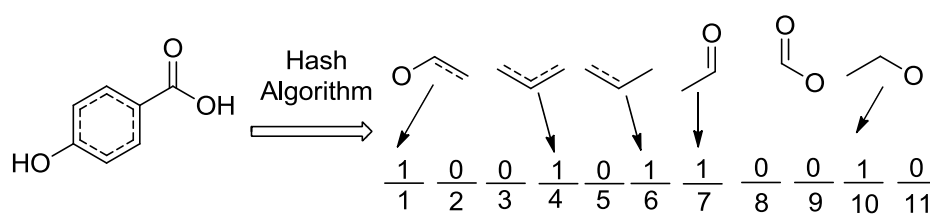


Figure 17 Molecular split and comparison with reference.

Fingerprint of molecule above: 1 0 0 1 0 1 1 0 0 1 0

Fingerprint of a query molecule (not shown): 1 0 1 0 1 0 1 0 1 1 0

The fingerprints for two molecules can be used to calculate structural similarity using the Tanimoto coefficient (or Jaccard coefficient). Tanimoto coefficient is commonly used to investigate structure similarity of compounds. $T(a, b)$ of molecules A and B is defined by:

$$T(a, b) = \frac{N_c}{N_a + N_b - N_c}$$

In this equation, N represents the number of attributes in each object (a, b), c in this case is the intersection set.

For Daylight fingerprints, two structures are considered dissimilar if $T < 0.85$.

$$\text{Therefore, } T_{\text{animoto}} = \frac{3}{(5+6)-3} = 0.375$$

Fingerprints are often used as a pre-screening tool in a substructure search to avoid running unnecessarily a graph isomorphism algorithm.

2.3.2.5 Shape and electrostatics

It is commonly recognized that shape and hydrogen-bonding capability are among the best description of a molecule that is independent from chemical structure if we want to understand molecular recognition properties (Böhm et al., 2004). Among commonly used tools, ShaEP superimposes molecules based on shape and electrostatic potential (Vainio, et al., 2009). BRUTUS is another rigid-body molecular superposition and similarity searching program based on grid algorithm, which also considers properties from charge distributions and van der Waals shapes of the compounds (Röckö et al., 2006). All these methods are essentially used for ligand-based screening through shape-based comparison.

2.3.2.6 Pharmacophoric representation

Pharmacophore is an old concept, which describes molecular features important for activity. These features include hydrogen bonding or accepting groups, hydrophobic or ionic centers, the distances that relate them and geometry in 3D space. Pharmacophoric representations can be used to compare molecules and they are mainly used as filters to query databases. An example of a pharmacophore search for protein-protein interaction is presented in Fig. 18. (Zhang, unpublished results)

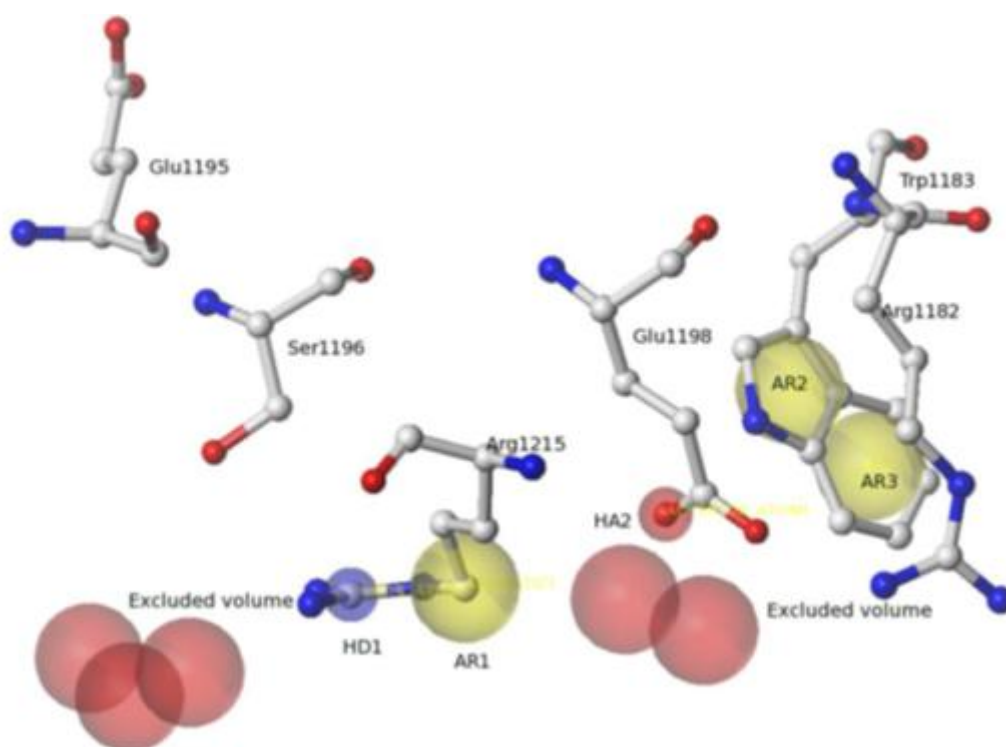


Figure 18 *An example of protein-protein interaction pharmacophore; Center, yellow lines in Yellow: aromatic (Ar), hydrophobic (Hy); Red: excluded volumes and hydrogen bond acceptors (HA), negatively charged groups (NC); Blue: hydrogen bond donors (HD), positively charged groups (PC)*

3 Aims of the study

The major aims of this study were to investigate small molecule mimics of phosphate groups, using both computational and synthetic approaches.

The specific research aims were as follows:

1. to construct a workflow useful to mine the Protein Data Bank for local structural replacements,
2. to identify previously unrecognized local structural replacements that could be useful bioisosteres of phosphate,
3. to uncover the local structural replacements of a ribose linker (3'-5' connected),
4. to characterize the structural mechanisms in the recognition of local structural replacements (bioisosteres),
5. to probe the human macrodomain protein 1 (MDO1) binding site with potentially bioisosteric replacements of the phosphate groups as well as distal ribose of ADPribose including ligand-based design, synthesis, and retrospective docking analysis, and
6. to discover compounds binding to MDO1 using a virtual screening approach.

4 Materials and Methods

4.1 Computational section

4.1.1 Material

- Protein Data Bank (PDB)

Since 1971, 98% of all structures of proteins or their complexes, mainly consisting of X-ray crystallography (88.6%) and NMR (9.4%) data have been stored in the PDB. One typical PDB format file contains atomic information of macromolecules, ligands, water and ions. The main content of PDB files is stored in the ATOM or HETATM rows. By tradition, ATOM is initial flag of macromolecules, such as proteins and nucleic acids, while HETATM remarks small molecules. Detailed information is then included, such as atom number, atom name, residue name, chain identifier, residue sequence number, atomic Cartesian coordinates, occupancy and temperature factor.

- The resolution filter

In X-ray crystallography, resolution is a measure of the quality of the fit of the modeled 3D structure with the electron density data. Ideally, X-rays are scattered identically if all protein molecules in a crystal are aligned in the same way. In reality, anisotropy of protein alignment in the crystal, several possible low energy conformations, and local flexibility results in a diffraction pattern that does not reflect the ideal case. The main difference of higher and low resolution is usually seen on surface loops and in long side chains (Arg, Lys, Glu, Gln) for which rotamers are might not be correctly modelled. The majority of the structures (93.6%) are within four ranges of resolution: 1-1.5 Å (9.0%), 1.5-2 Å (38.1%), 2-2.5 Å (30.6%), and 2.5-3 Å (15.9%). Hydrogen atoms are not seen unless for the highest quality structures with resolution of 1 Å or under (Fig.19A). With resolution of 3 Å or lower (Fig.19D), we can only see the contours of protein chain while the position of side chain is less accurate. In our study, we used 2.8 Å as a resolution filter. We suppose this criterion is suitable to allow us obtain enough relevant data. The data is meant to be

analyzed by chemists, who themselves can exert caution when knowing that a given example is of relatively low resolution.

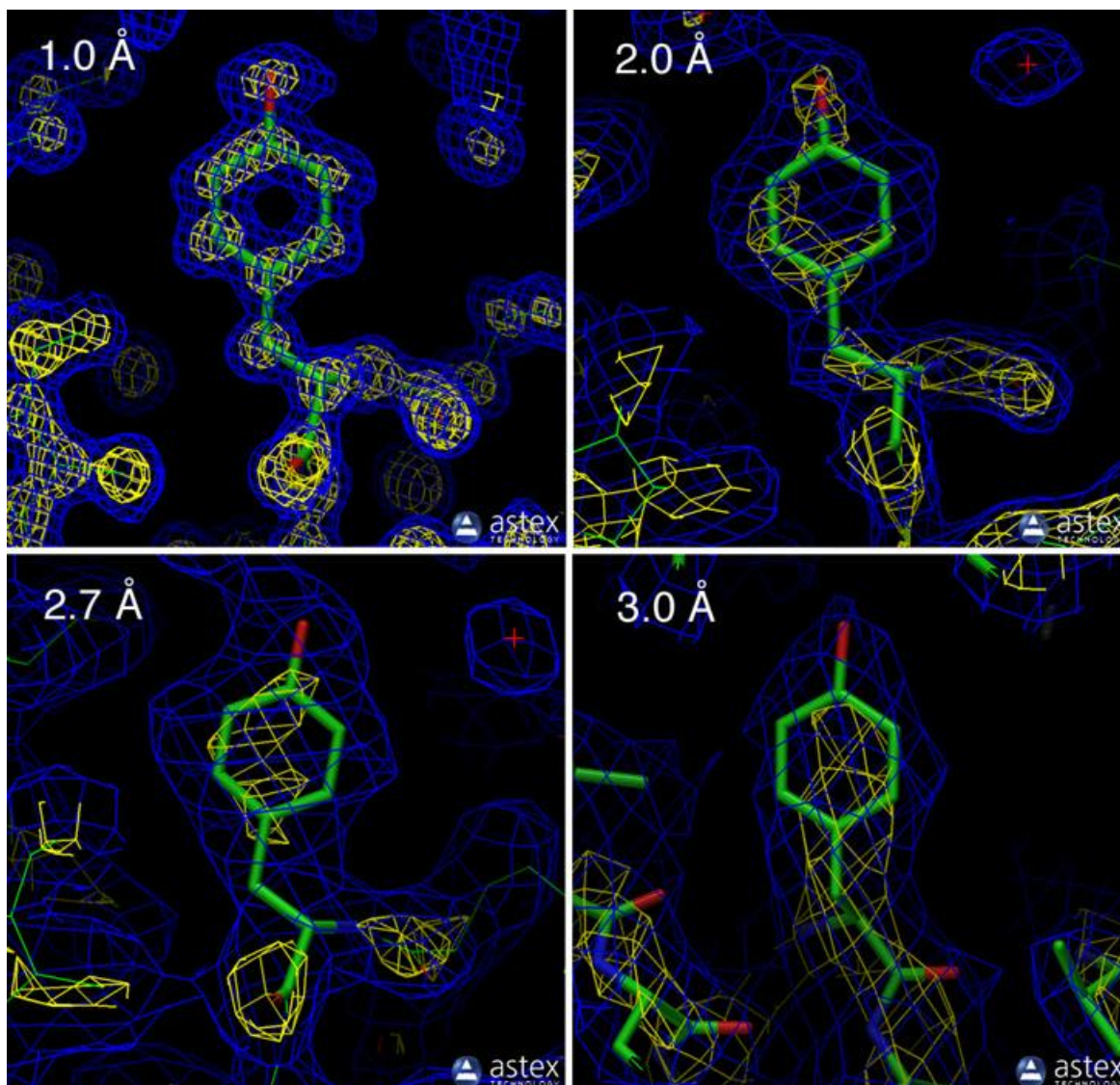


Figure 19 *Electron density maps for structures with a range of resolutions are shown. The first three show tyrosine 103 from myoglobin, from entries 1A6M (1.0 Å resolution), 1O6M (2.0 Å resolution), and 1O8M (2.7 Å resolution). The final example shows tyrosine 130 from hemoglobin (chain B), from entry 1S0H (3.0 Å resolution). In the pictures, the blue and yellow contours surround regions of high electron density, and the atomic model is shown with sticks. Reproduced with permission of the PDB. (Berman, 2000)*

4.1.2 Methods

- Data mining

In this work, a computational automated workflow written in Python language was built (Fig.20).

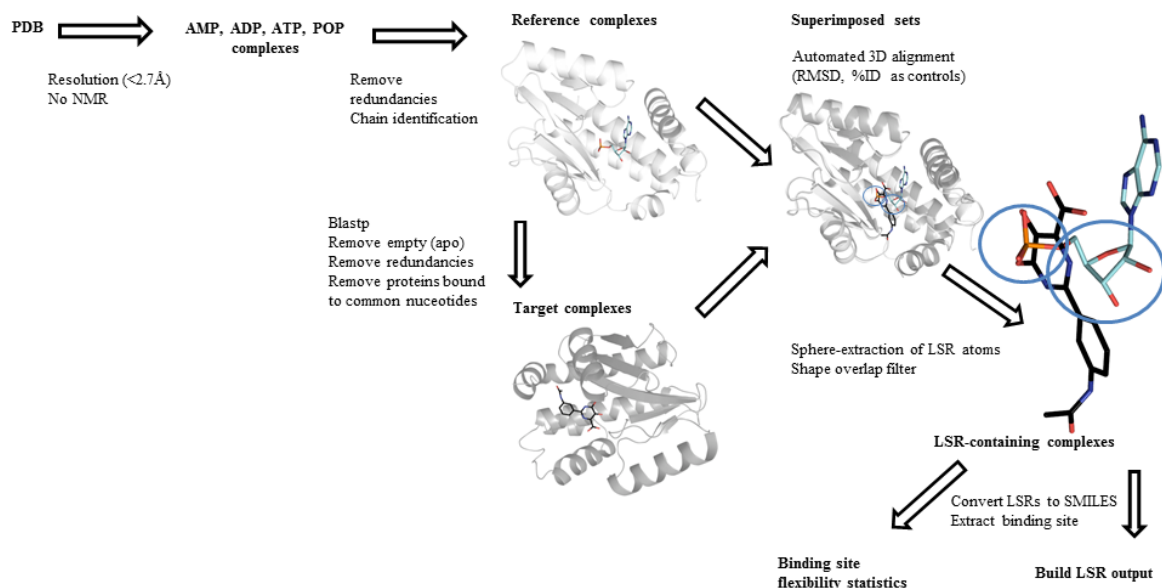


Figure 20 Flow chart of data mining.

The workflow is applied to extract the local structural replacements (LSR) that occupy the positions of the individual α , β and γ phosphates or ribose moieties in POP, AMP, ADP and ATP-binding proteins. The workflow can be parameterized by many aspects and the individual parameters used for the run presented are given at the end of this section.

Technically, the workflow uses and connects different external tools: (1) Blastp for sequence-based retrieval of protein homologs (Johnson et al., 2008), (2) TM-align for protein structure superimposition (Zhang and Skolnick, 2005), (3) ShaEP for assessing a score for the ligand-based superimposition of molecular fragments (Vainio et al., 2009), (4) Babel to convert extracted 3D structure files to SMILES code (O'Boyle et al., 2011), (5) Needle from the Emboss package to conduct global sequence alignment (Rice, et al., 2000). Setting up the workflow required developing extra lines of code for example the connectivity needed to be rewritten in order to run ShaEP properly and manage missing or incorrect connectivity in PDB files. For the purpose of visualization and data analysis, other python tools have been used, for example a tool that implemented the Kabsch algorithm to superimpose small molecule fragments (Borrel et al., unpublished). The software R was used for all statistical analysis and presentation of data.

At the first step, the workflow retrieve from PDB (Berman, 2000) the three-dimensional structures of proteins that are homologous to proteins with a bound POP, AMP, ADP, and ATP (named “reference ligands”; reference ligands are each bound to several “reference proteins”). Protein chains are often problematic for automated extraction protocols and a single reference ligand is taken for any reference PDB files. Proximity (at least one atom within 4.5 Å) is used to select the protein chain(s) of interest, and a global sequence alignment used to eliminate redundancy in the protein chains of a given PDB file. The protein chain(s) considered are (is) appended to the PDB file name in the output of the workflow.

Secondly, for each reference protein found in the previous step, the workflow retrieve and superimpose homologues and keep those with a non-nucleotide ligand bound at equivalent position to the reference ligand (these homologs are referred to as LSR-containing ligand and LSR-containing protein). Chains of the LSR-containing proteins are identified by proximity (4.5 Å) as above and appended to the file names. Metrics (RMSD, length of superimposed region and percent identity) are output from the whole body protein superimposition. In this study, only near identical proteins are considered and perhaps therefore no global RMSD deviation above 3.5 Å was found and all superimposition were kept in the study (a high RMSD indicates a poor superposition that may be due to conformational differences such as motions upon ligand binding due to protein being homologues). At that stage, data on the binding site flexibility (C_α and all atoms RMSD; longest deviation) is extracted for the subset of proteins with exactly identical amino acid type and numbers in the reference and in the LSR-containing sets. A set of overrepresented ligands are automatically removed to speed up the analysis, based on a manually constructed list; this list is a parameter in the workflow.

In the third step, LSRs are extracted at the level of individual groups; for example ADP is composed of three “replaceable” groups that we study here, i.e. ribose, α -phosphate, and β - phosphate. In the case of phosphate, extraction of atoms belonging to the LSR is based on a single sphere centred on the phosphate; and in the case of ribose extraction is based on several spheres centred on every atoms (O5', C5', O4', C4', O3', C3', O2', C2' and C1'). Target proteins with empty LSRs are removed. Redundant groups are then merged and some complexes filtered out when they do not conform to set criteria, for

example a too small overlap of the LSR on top of the considered group of the reference protein (α -, β - and γ - phosphate or ribose). The SMILES of the extracted local structural replacements are then used to sequentially assign to the “replacing” groups exclusively one of nine non-overlapping categories containing: phosphorus, cycle, SO_2 ; CON (generally carbamoyl or amide); COO (ester or carboxylic acid); exclusively C; only C and O; only C and N; only C, O and N; Other.

The empirically optimized filters used for the run, which is presented are as follow: 1) PDB August 2014 release (102158 entries); NMR entries are not considered; 2) resolutions are better than 2.7 Å; 3) homologues were found by a BLAST e-value filter of 10^{-100} ; 4) manually built list of pre-filtered compounds: TTP, DCP, DGT, DTP, DUP, ACP, AD9, NAD, AGS, APC, AOV, UDP, GDP, GTP and CTP, in addition to AMP, ADP, ATP and POP (in the version of the paper submitted, ANP will be further removed); 5) definition of the binding site as any amino acid with at least one atom within 4.5 Å of the bound reference ligand; sphere extraction of atoms belonging to the LSR in the case of phosphate with a radius of 3.0 Å; and in the case of ribose of 2.5 Å; a shape component of the score computed by ShaEP larger than 0.2 required.

- **Python scripting**

The workflow is developed in Python. Guido van Rossum created the Python in 1989. Python is probably the most popular programming language, and it is used not only by IT giant such as Google, Yahoo but NASA. Python is multi-paradigm: procedural, declarative, functional, and object-oriented. So far, two basic versions of Python were released: 2.0 on 16th October 2000, Python 3.0 on 3rd December 2008. The latest Python versions are 2.7.8 and 3.4.2. When this project started, I used Python 2.7.2. The 2.x and 3.x series of Python will work for most cases, with a few minor changes, for example “*print*” is a statement in Python 2.x instead of a function in Python 3.x.

Python has a large and comprehensive standard library. The standard library contains several kinds of components: data types, built-in functions and modules. Besides, there are several extensions such as Biopython (Cock et al., 2009) a tool for biological computations; Numpy (Walt, et al., 2011) providing a static array data type, which speed-up numeric computations; Matplotlib (Hunter, 2007) 2D and 3D plotting library for

Python; and Scipy based on Numpy, to do scientific computing, such as statistics, optimization, and Fourier transforms.

In this chemoinformatics project that was not computationally nor graphically demanding, the python standard library was used, while the figures were plotted with R (R Development Core Team, 2011). R is a software facilities suite for data manipulation, calculation and graphical display.

- **BlastP**

Basic local alignment search tool (Blast) was used to retrieve all the structures of a given (reference) protein sequence. Blastp uses a heuristic algorithm for comparing primary protein amino acid and DNA nucleotide sequences expressed as a string (a succession of letters, 20 for proteins and four for DNA). (Altschul et al., 1990) One of the most widely used bioinformatics program for sequence searching, a BLAST search allows comparing a query sequence with a sequence database, and identifies target sequences that resemble the query sequence above a certain threshold. BLAST use high-scoring segment pair (HSP) scores to characterize the match of query sequence and individual one in database. The significance of this match, expressed as expectation or expects value (E-value), relates to the number of hits of a given match score one can "expect" to see given the size of the database and the length of the query. The E-value decreases exponentially as match score increases.

In this study, the e-value setting-up of protein-protein BLAST "blastp" the threshold for retrieving proteins identical to queries available in the PDB was empirically set to 1×10^{-100} . This value should allow to recover mutants, truncated proteins (e.g. fragments not seen in crystallography), or proteins whose sequence contain a purification tag such as the His-tag.

- **TM-align**

TM-align is a software that identifies the optimal alignment between the tertiary structure of protein pairs (Zhang and Skolnick, 2005). TM-align gives the sequence identity value (ID) calculated over the C-alpha carbons that are used to make the alignment of structures.

The Root Mean Squared Deviation (RMSD) of protein pairs is also provided. The RMSD is defined given two sets of n point of graph **v** and **w**:

$$\begin{aligned}
RMSD(\mathbf{v}, \mathbf{w}) &= \sqrt{\frac{1}{n} \sum_{i=1}^n ||\mathbf{v}_i - \mathbf{w}_i||^2} \\
&= \sqrt{\frac{1}{n} \sum_{i=1}^n ((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)}
\end{aligned}$$

Superimpositions of paired proteins were done by translation vector and TM-score rotation matrix. The matrix and vector are obtained from the coordinates of all counterpart atoms in paired proteins. I embedded TM-align into Python scripts to automate the proteins superimposition.

- **Finding reference atoms**

It was technical difficult to locate the atoms belonging to phosphate or ribose fragments to be used as references in the PDB file. Here, this could be done taking advantage of the labels in the PDB files. If not for these labels, I would have been required to use substructure search algorithm such as the Ullmann algorithm, (Ullmann, 1976), which is less straightforward than it seems because the connectivity matrix of HETAM is not found in PDB files.

In a preliminary version of the workflow, the bridge oxygen of POP and ADP dataset was assigned as centroid of the probe sphere. For the AMP dataset, the center atom was a phosphorus atom. For ribose, a sphere centroid (a *pseudoatom*) was defined using the center of mass of the extracted coordinates of O4', O3' and O2' atom. Starting with the centroid atom coordinates obtained from a reference protein, the search was terminated if there were less than 3 atoms within the probe sphere with radius 3 Å. Otherwise, the sphere probe was expanded up to 4.5 Å. Atoms of ligands co-crystallized with superimposed protein within this space were extracted, defining as structural replacement of phosphate or ribose.

During the analysis of the preliminary result, we thought it would be more accurate if the center atom was assigned to each individual phosphorus atom. This idea was also applied to the ATP dataset that was added in this second version. In addition, we used a shape-based superimposition to control the relevance of the extracted fragment (software ShaEP,

see below).

- **ShaEP**

ShaEP, developed by the group of Mark S. Johnson from Åbo Akademi University (Vainio, et al., 2009), was used to produce volume- and electrostatic- overlap (shape) scores. ShaEP contains two main functions: a scoring function, that scores the superimposition of atoms represented by their electrostatic potential and local shape using Gaussian functions; and a genetic algorithm that finds the superimposition that maximize the score overlap of the molecules. In my study, only the scoring was conducted since further optimization may change the extracted substructure spatial characteristics from protein-ligand complex.

4.1.3 Homology modeling, docking and virtual screening

- **Molecular modeling of MDO1**

A structural model of the ligand-bound conformation of human MDO1 was constructed based on the crystal structure of *Archaeoglobus fulgidus* Af1521 (PDB: 2BFQ)(Karras et al., 2005) using the software Swissmodeller (Biasini et al. 2014). The modeling was done early in the study, before the crystal structure of MDO2 became available, but this should not affect the conclusions since the ADPr binding site is well predicted and likely not influenced by the template. Before docking, the modeled human MDO1 was handled with protein preparation wizard (Schrödinger Ltd). The atom type, side chain protonation states, and hydrogen network in the protein were assigned using PROPKA at pH 7.0. Glide receptor grids were generated by defining a 10 Å box localized at the centroid of ADPr copied from the *A. fulgidus* macrodomain protein (PDB: 2BFQ).

- **Ligand preparation**

Two tools were used: Balloon, for the study conducted in 2010-2011; and Ligprep (Schrödinger Ltd), for which the license was acquired later, for the studies conducted in 2012-2014. Balloon, free software from the Mark S. Johnson group, is enumerating low energy 3D conformations of molecules from a 2D format, using a genetic algorithm and a Pareto multi-objective evaluation (Vainio et al., 2007). The input format can be MOL2, SDF or SMILES. Balloon performs energy minimization using the MMFF94-like force

field. By default, Balloon uses distance geometry only.

For the analysis of the binding modes of the synthesized compounds to MDO1, the possible ionization states of ligand candidates were generated by LigPrep Suite with default parameters. The global minimum energy conformation of each structure was identified using OPLS 2005 force field and no tautomers were generated.

- **Docking simulations**

Molecular docking, as a routinely used computational method, helps to model molecular recognition events more rapidly and at lower costs than X-ray crystallography (but at the expense of accuracy). Docking programs try to simulate the complementary relationship between ligand and protein in terms of volume and interaction. Molecular docking address three questions: 1) finding the native binding pose of individual ligands in protein binding site; 2) estimate the relative binding affinity of each individual ligand and 3) discriminate ligands from non-binders (decoys). Technically, the docking problem is divided into a sampling problem and a scoring problem.

The sampling process should fish out experimental binding pose of ligand in binding site and scoring functions should give rise to highest score among all binding poses. So far, docking software routinely includes ligand flexibility. Allowing the flexibility of the protein active site, at least at the level of side chains, is possible but extremely expensive computationally, and leads to difficult scoring problems. Complete flexibility of the system or explicit inclusion of water molecules is beyond reach of most computer resources (Meng et al., 2011).

Scoring functions are divided into three types: force-field-based, empirical and knowledge-based. Traditional force-field-based estimates the binding affinity by summarizing the non-bonded (electrostatics and van der Waals) interactions between sampled ligand and protein pose. Advanced force-field-based scoring functions also include hydrogen bond, solvation and hydrophobic contribution. Knowledge-based functions extract structural information from Protein Data Bank (PDB) and Cambridge Structural Database (CSD), and employ the Boltzmann law to transform the atom pair preferences into distance-dependent pairwise potentials (Shen et al., 2011). “Empirical”

scoring functions use a training set containing binding affinities.

During the course of my thesis, I have been using four docking programs: Surflex-Dock (Tripos) (Chikhi, et al., 2008), Gold (Hartshorn et al., 2007), Autodock (Morris et al., 2009), as well as Glide-SP and Glide-XP (Schrödinger) (Friesner et al., 2006). Gold was the main software used by the group in 2010-2012 until the Schrödinger software license was acquired.

For the analysis of the binding modes of the synthesized compounds, to MDO1, the molecular docking experiments were carried out with Maestro suite Glide (Friesner et al., 2006). Default parameters were used. Van der Waals scale factor was 1.0, and partial charge cutoff was 0.25. Ligands were flexibly docked into the binding groove using the Glide XP protocol.

4.2 Synthetic chemistry section

- **Materials and general procedures**

All reagents were commercially available. They were acquired from Fluka (Buchs, Switzerland) and Sigma-Aldrich (Schnelldorf, Germany) and were used without further purification. All reactions in anhydrous solvents were conducted in oven-dried glassware under anhydrous argon. Thin-layer chromatography (TLC) was performed using Silica Gel 60 F₂₅₄ (Merck) and Silica Gel 60 NH₂ F_{254s} aluminium sheets (Merck), visualized by UV illumination and stained with ninhydrin in EtOH (1.5% w/v). The ¹H and ¹³C NMR spectra were recorded on a Varian Mercury-VX 300 spectrometer as solutions in CDCl₃, DMSO-*d*₆ and CD₃OD. Chemical shifts (δ) are reported as parts per million (ppm) relative to the solvent peak (CDCl₃ 7.26 and 77.21 ppm, DMSO-*d*₆ 2.50 and 39.52 ppm, CD₃OD 3.31 and 49.00 ppm). Multiplicities of peaks are represented by s (singlet), d (doublet), t (triplet), q (quartet), qn (quintet), and m (multiplet). HPLC–MS analyses were performed to determine purity of all tested compounds. Purity of all tested compounds was > 95%. Mass spectra were measured on a Bruker Daltonics Esquire-HPLC spectrometer, with XTerra MS RP18 column (4.6 mm × 30 mm, 2.5 μ m) or on a JEOL JMS-AX505 (Tokyo, Japan) spectrometer with direct input and electron ionization (EI).

4.3 Biological testing

This section introduces experiments conducted by our collaborators. ITC, DSC and DSF experiments were conducted at the University of Oulu (Finland) in the group of Docent Lari Lehtiö. ADPribose and poly-ADPribose radiolabeled competition assay were conducted in the group of Dr. Päävi Tammela, University of Helsinki, Finland

- **Competition binding assay**

Binding assays aimed to measure interactions between two molecules, such as a protein binding into small molecule. Competitive binding assays were based on competition of unlabeled ligand (passive) against labeled one (active) during protein-ligand reaction. The common protocol of a competition assay was performed such that the binding site of receptor was saturated by radiolabeled ligand firstly, then doses of the unlabeled one were added and displacement of the labeled ligand measured. This experiment was repeated for different doses of the unlabeled ligand. IC_{50} , the concentration of unlabeled ligand necessary to displace 50% of the labeled one, can then be estimated. K_i , which was independent of the labeled ligand binding affinity, can then be estimated from IC_{50} using for example the Cheng-Prusoff equation.

Competitive assay for poly-ADP-ribose (poly-ADPr) binding. MDO1, expressed in *Escherichia coli*, was coated on yellow Delfia 96-well plates (PerkinElmer) and unbound protein was removed by washing after overnight incubation. Biotinylated poly-ADPr, produced by recombinant human PARP1 from a mixture of NAD^+ and 6-biotin- NAD^+ , was added to the wells containing MDO1 and incubated for 1 h at RT on a shaker. Unbound poly-ADPr was removed by washing with $3 \times 600 \mu\text{L}$ of 10 mM MOPS (1 mM NaCl, pH 7) by using a plate washer. Europium-labelled streptavidin (40 ng/mL in Delfia Assay Buffer) was added and after incubating at RT for 30 min, unbound streptavidin was removed by washing ($4 \times 600 \mu\text{L}$, Delfia Wash Buffer). The plate is left to dry at RT and then 200 μL of Delfia Enhancement solution was added to the wells. After 5 min shaking at RT, time-resolved fluorescence at excitation 340 nm, emission 615 nm was measured with a Victor plate reader. The ability of compounds to compete with poly-ADPr binding to MDO1 was assessed by comparing the europium signals from treated and non-treated wells.

Competitive assay for ADP-ribose (ADPr) binding. MDO1 was coated on Scintiplates (PerkinElmer) as described above. ^3H -ADPr was added and after incubation, unbound ^3H -ADPr was removed by washing. The plate was dried at RT and measured by Microbeta scintillation counter (PerkinElmer). The ability of compounds to compete with ^3H -ADPr binding to MDO1 was assessed by comparing the signals from treated and non-treated wells.

- **Isothermal titration calorimetry**

Isothermal titration calorimetry (ITC) was used to determine how much energy was released when protein-ligand binding takes place, providing an estimate about the enthalpy change (ΔH). ITC used thermocouple circuits to detect temperature differences between the reference cell and the sample cell with the receptor. Ligand was added into the sample cell, causing heat either absorption or release. The power needed to maintain balanced temperatures between the sample and reference cells were measured.

ITC also provides a quantitative measure of the binding affinity (K_d) of a ligand to a protein. According to relationship $\Delta G = -RT \ln \left(\frac{K_d}{c^\theta} \right) = \Delta H - T\Delta S$, Gibbs energy ΔG and entropy change ΔS can be then calculated.

- **Biological assays of MDO1 using differential scanning fluorimetry**

Differential scanning fluorimetry (DSF) is used to investigate to which extent ligand binding could stabilize the protein structure. Protein unfolding happens with a certain temperature range, the midpoint of the transition (defined as T_m) from protein folding to unfolding is different in absence and presence of ligands. Increased transition temperature when a compound is present indicates binding; the higher T_m implies the tighter ligand binding. Fluorescence-based thermal shift assay uses a dye that is fluorescent in hydrophobic environment but not aqueous solution to characterize the ligand binding to protein. Upon temperature-triggered protein unfolding, the protein hydrophobic core becomes exposed to the dye and triggers the fluorescence.

Testing protocol: The expression construct for human MDO1 (residues 58-325) with an N-terminal 6×His-tag and a thioredoxin fusion tag followed by a TEV cleavage site (pNH-TrxT) was a generous gift from the Structural Genomics Consortium, Oxford. Protein was expressed by transforming the plasmid into *Escherichia coli* Rosetta2 (DE3) cells. Glycerol stock was used to inoculate small 5-mL pre cultures grown overnight in Luria broth (LB) media, which were then used to inoculate two large 1-L cultures containing Terrific broth (TB) auto induction media with trace elements (Formedium, UK). The media was additionally supplemented with 8 g/L glycerol. The cultures were allowed to grow up to OD₆₀₀ of 1.5 at 37 °C with shaking after which the temperature was reduced to 18 °C and the cultures were allowed to grow overnight (14-16 h). The cells were then collected by centrifugation (5500×g, 25 min, 4 °C), resuspended in lysis buffer (50 mM HEPES, pH 7.5, 500 mM NaCl, 10% glycerol, 10 mM imidazole, and 0.5 mM TCEP; 1.5 mL/g of cells) and stored at -20 °C.

The cell suspension was thawed, 10 µg/mL DNAase, 0.1 mM Pefabloc, and 0.25 mg lysozyme was added followed by incubation on ice for 5 min. The cells were disrupted by sonication and the lysate was cleared by centrifugation (35000×g, 30 min, 4 °C). The supernatant was filtered using a 0.45 µm filter syringe before it was loaded to two 1-mL HisTrap HP columns (GE Healthcare, UK), which were pre-charged with Ni²⁺ and pre-equilibrated with binding buffer (20 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 10 mM imidazole, and 0.5 mM TCEP). The bound protein was washed with 5 column volumes (10 mL) of buffer supplemented with 25 mM imidazole and eluted with buffer containing 350 mM imidazole. The eluted protein fraction was concentrated and diluted in size exclusion buffer (20 mM HEPES, pH 7.5, 250 mM NaCl, 5% glycerol, and 1 mM TCEP) to a final volume of 10 mL to bring the imidazole concentration down to 50 mM. TEV protease was added (1:30 molar ratio) and the sample was incubated overnight at 4 °C. The cleaved protein was run again through a 1-mL HisTrap HP column (GE Healthcare, UK). The flow-through was collected and concentrated to perform size exclusion chromatography using Superdex S-75 column (GE Healthcare, UK). The collected fractions were pooled, concentrated and divided into small aliquots which were flash frozen in liquid nitrogen and stored at -80 °C.

The binding of the compounds was tested using DSF. The concentration of the protein used was 0.25 mg/mL in a reaction volume of 30 µL. Sypro orange was used as a reporter

dye. The different ligands that were tested were used at a final concentration of 1 mM. All the conditions were performed in triplicates. MDO1 without any ligand and MDO1 with a 1 mM AMP, ADP and ADPr were used as controls. The experiment was performed on a RT-PCR machine with the starting temperature of 20 °C after which the temperature was increased at 1 °C for every 60 seconds until the temperature reached 90 °C. The fluorescence readings from the environmentally sensitive Sypro orange reporter dye were recorded for every 1 °C increment of the temperature. Data were normalized and analyzed using Boltzmann sigmoidal curve fitting GraphPad Prism 5.0 (GraphPad software Inc., US).

5 Results and discussion

For clarity, the compounds synthesized as part of the thesis work are numbered starting from **1**. The results presented in this section follow the specific aims 1-6.

In this thesis, I first computationally studied the local structural replacements of phosphate. These replacements are not “linked” to biological activities, but only to observation of molecular complexes in PDB structures. Nevertheless, these local structural replacements should provide medicinal chemists with new ideas for synthesis of novel compounds. Furthermore, many of the compounds shown as examples have nanomolar or micromolar activities, which suggest that these replacements are favorably binding to the protein. As a result, we recognized recurring mechanisms that will enrich our understanding of the molecular recognition of phosphate groups. I then designed two series of ADPr analogs that contained replacements of the phosphate groups, and synthesized, purified and analyzed them. Five active compounds were found to bind to the MDO1 protein when tested using DSF. The virtual screening results will be presented at the end of this section.

The most critical concepts in this thesis are those of *local structural replacements* and of *bioisosteres*. These concepts are overlapping but not identical.

- *Bioisosteres* emphasize similarity in chemical properties – isosteres – as well as in biological activities, i.e. similar binding affinities. Nonetheless, biologically active compounds with bioisosteric replacements may be the result of completely or partially different modes of binding.
- *Local structural replacements*, on the other hand, contain the idea that the fragments studied occupy the same location in the binding site. Nonetheless, biological activity (binding affinities) may vary greatly. Furthermore, the binding affinity depends on the chemical structure of the complete ligand, not only the local structural replacement considered. Molecular interactions may or may not be conserved when local structural replacements are used, as will be presented in this thesis. None of the local structural replacement presented here is “new” in the chemical sense – all have been synthesized and crystallized in order to be found in

the PDB. Therefore, it is the comparison with a reference protein and the recurring aspect (i.e. the same replacement is found several times) that makes them novel or interesting.

5.1 Workflow to mine the Protein Data Bank for local structural replacements

- **Data collection**

A summary of the data extracted at each step of the workflow is presented in Table 4.

Table 4. *Number of LSR-containing complexes and ligands collected in each of the four datasets*

	Reference complexes in PDB	Reference complexes filter	LSR-containing + empty		LSR-containing complexes (empty eliminated)	Unique LSR-containing ligands	After ShaEP filter			
			Total	with BS identical to BS reference			Ribose	α -P	β -P	γ -P
AMP	430	116	1181	358	576	409	227	205	-	-
ADP	1372	314	3576	1390	1050	703	572	523	400	-
ATP	814	271	3067	1048	1006	730	461	407	318	239
POP	202	42	286	118	240	121	-	51	59	-

We constructed a total of four sets of 116, 314, 271, and 42 unique proteins that are in complex with pyrophosphate (POP), AMP, ADP, or ATP. Each of these reference proteins is also crystallized in complex with at least one LSR-containing ligand. The workflow is automated and in the run presented, criteria were empirically selected for optimal parameters that avoid depleting the dataset from otherwise interesting examples while leaving decision as to whether complexes are interesting to the end-user. The LSR-containing proteins are first selected using a resolution filter of 2.7 Å, and a criteria of 10^{-100} expect value to the Blastp search. This second criteria was chosen so that it “let through” very related proteins, which may come at valuable examples. These may be closely related species or proteins that have been mutated or engineered (at the binding site or elsewhere, for example by addition of a His-tag for purification). The selected proteins are in ~30%

(2914/8468) of the cases entirely identical at the binding site (same of amino acids, same numbering) (Table 4).

The LSR-containing proteins were superimposed on top of their respective references (Figure 21). As a result, the superimposition appears very reliable, which is not surprising considering that identical or near identical structures are at hand. During the rather exhaustive manual analysis of the data, we did not find local structural replacements that were offset as a result of erroneous or ambiguous whole-protein superimposition; even when conformational changes did take place they were limited to a restricted region of the binding site.

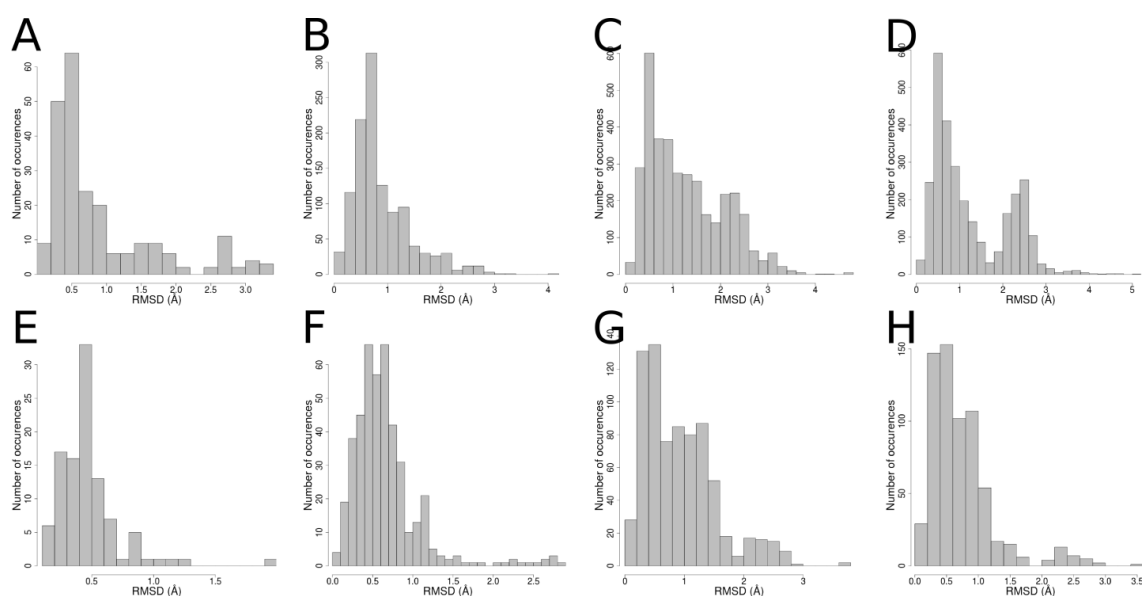


Figure 21 *Distribution of for protein superimposition (A-D), RMSD of completely identified protein with reference (Å); (E-H), all target protein RMSD (Å). RMSD are for the datasets (A, E) POP; (B, F) AMP; (C, G) ADP and (D, H) ATP datasets.*

In total and after elimination of redundancy and empty binding sites, 929 unique LSR-containing ligands were found, that when taking together their diverse occurrences they account for 3462 local structural replacements. The most represented LSR-containing ligands are adenosine derivatives (PDB ligand code ANP, 2267 occurrences; 3P1, 151; IMP, 115), kinases inhibitors (STU, 183; L20, 123; R68, 91; R69, 99; VX6, 83), HSP90 inhibitors (3FD, 147; GDM, 143), and crystallization agents bound at the nucleotide site (BET, 149; SEP, 99).

- **Reference groups**

The 116, 314, 271 and 42 sets of reference proteins represent a variety of folds. Kinases proteins account for a majority of the reference proteins: kinase 28% (205 out of 743), HSP 5% (34) and others 68% (504). It is not surprising given the large size of the kinome and their key role as anticancer drug targets. The other proteins under study are Heat Shock proteins, proteins active on DNA and RNA such as ligase and synthases, and proteins involved in metabolic reactions, as well as myosin.

All these proteins are bound to nucleotides, which presumably represent their natural endogenous ligands. The terminal α -, β - and γ -phosphates of respectively AMP, ADP and ATP, as well as both phosphates of POP, contains three oxygen atoms not bonded to heavy atoms: two hydroxyl groups that may play roles as hydrogen bond donor/acceptors, and a carbonyl group that may be a single hydrogen bond acceptor. Upon ionization, the hydroxyl groups act as a hydrogen-bond acceptor and individually may carry one negative charge; the terminal phosphates therefore may carry up to two negative charges. Connected internal phosphate groups, i.e. the α -phosphate of ADP and both the α - and β - phosphates of ATP contain one free hydroxyl group and therefore may be singly negatively charged. At physiological pH, for example the AMP phosphate group is negatively charged with a first pK_{a1} of 2.12 and a second pK_{a2} of 7.21 (W.D. Kumler; John J. Eiler, 1943).

In the collected reference proteins, neutralization of the charges by coordination of a metal appears to be an important factor in the binding mode of phosphate oxygen atoms. Magnesium is the metal preferably found in the reference proteins close to POP, ADP and ATP, in many cases sandwiched in between oxygens (Figure 22). For AMP, neighboring metals differs and there is less metal coordination, which may reflect the lack of coordination of the metal by two phosphate atoms (Table 5). Phosphate binding has been shown to occur preferentially at phosphate binding cups, which happens often at the termini of α -helices and takes advantage of helical macrodipoles (Denesyuk & Denessiouk, 2003).

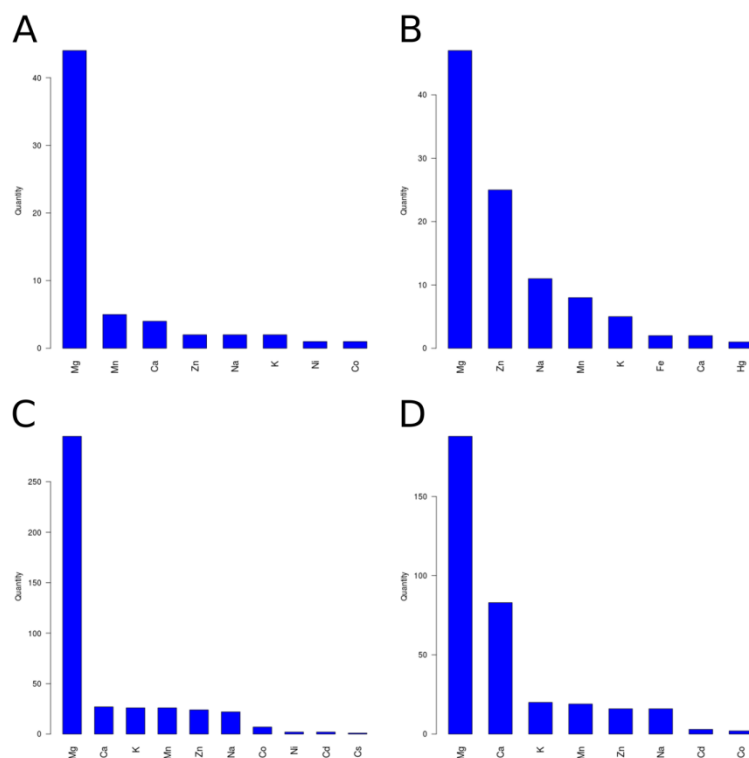


Figure 22. *Distribution of metals less than 3 Å of any oxygen atom of (A) POP, (B) AMP, (C) ADP, and (D) ATP.*

Table 5. *Number of metal atoms in the binding sites less than 3 Å from nucleotide oxygen atom; in the reference proteins only.*

	Reference protein			
	AMP	ADP	POP	ATP
Metals shared by α/β or β/γ (several metals for each complex possible)		147	57	228 (α/β -P) 113 (β/γ -P)
1 metal in binding site	16	247	21	180
2 metals in binding site	30	59	18	53
3 and more metals in binding site	1	23	12	34
Total with at least one metal	47	329	51	267
Total reference proteins considered (dataset studied less filtered than the final reference set)	189	488	92	374

The spatial location of phosphate groups relative to nucleotide base show that the β -pyranose form is likely to be the most predominant, as can be inferred from the Figure 23. The spatial location of phosphate groups relative to nucleotide base show that the β -

pyranose form is likely to be the most predominant, as can be inferred from the Figure 23. The distance between O3' and the closest oxygen atom bound to the α -phosphate group suggest that hydrogen bonding between these atoms may occur in a small proportion of AMPs (38 out of 282 distance between these heavy atoms less than 3.2 Å) (Fig. 24); this is not the case for ADP and ATP and need for further hydrogen bonding may result from the relatively smaller number of metals.

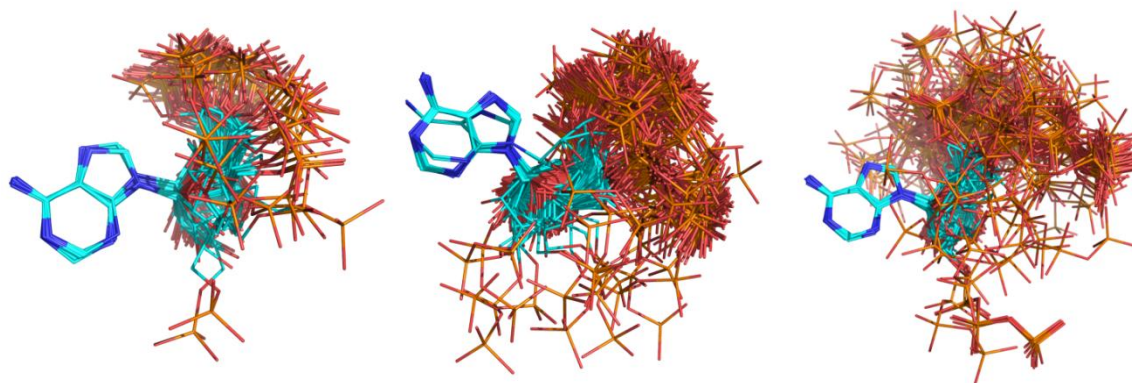


Figure 23. *Superposition of AMP (n= 189), ADP (n=488) and ATP (n=374) in the reference structures.*

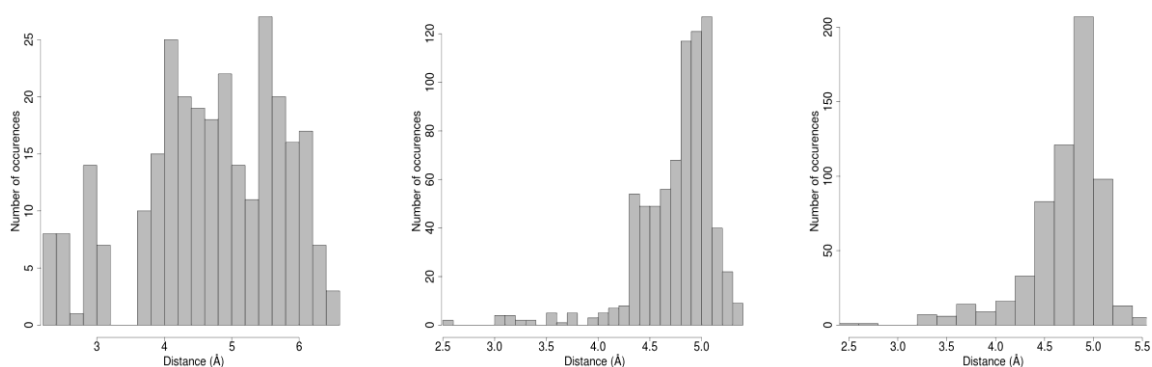


Figure 24. *Distribution of the minimal distances between O3' and the closest oxygen atom of O1P, O2P or O3P. Data is for (A) AMP; (B) ADP; (C) ATP.*

- **Local structural replacements**

929 different ligands were extracted using the workflow and data about the composition of the different groups is shown in Table 1. LSRs were then extracted and, in order to facilitate the analysis and to limit the size of the output, divided into non-overlapping groups based on the SMILE they contain (Fig.25).

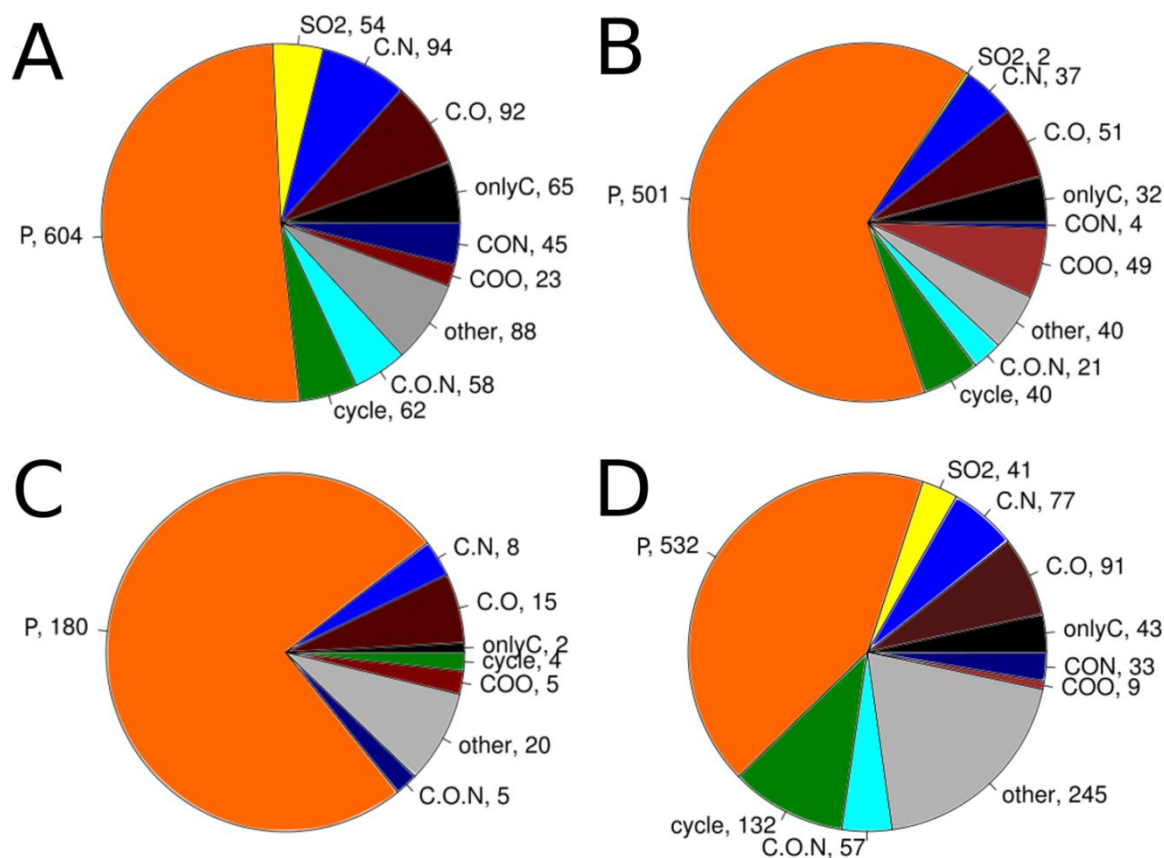


Figure 25. Subdivision of the dataset into categories assigned according to the SMILES codes of the local structural replacements; a single category is always assigned even when several are possible. (A) replacements of α -phosphate of POP, AMP, ADP and ATP; (B) replacements of β -phosphate of POP, ADP and ATP, (C) replacements of γ -phosphate of ATP; and (D) replacements of ribose of AMP, ADP and ATP.

We sequentially attributed ligands to categories, in order: P, presence of letter P in the SMILE code, cycle, presence of the number 1 which indicates a cycle, SO₂, COO and CON, any SMILES compatible with S or C connected to two O, or C connected to O and N. The remaining ligands were assigned to four groups, only C, O and N; only C and O, only C and N; only C; and what remained. The name of the categories in practice corresponds to *several* SMILES codes, i.e. alternative ways to connect a C atom with two O atoms are covered by our methods.

- **Phosphate structural replacements**

Structural replacements found in the context of the ligands are showed in Fig.26. It should be noted that the SMILES based classification is exclusive, i.e. the same type of replacements can be found in several categories.

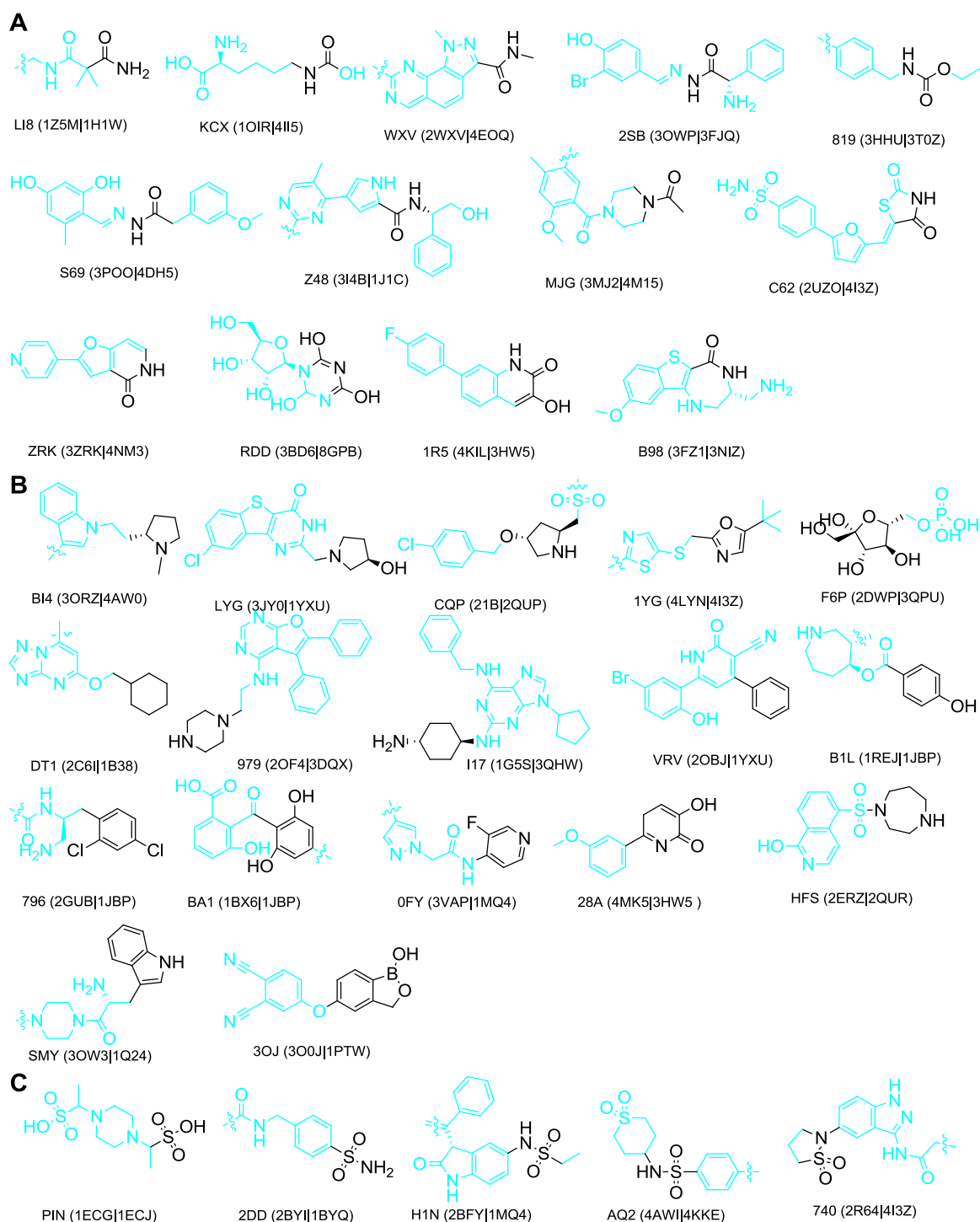


Figure 26. The exemplified local structural replacements of phosphate (black) (A) CON (B) cycle (C) SO₂. All examples are labeled as following ways: ligand three letters or digital code (code of protein who complex with the ligand| code of reference protein).

Phosphorus (P letter) containing SMILES are the most abundant in the sets. Many ligands are phosphate containing: nucleotide derivatives (e.g. ligand NAI, LSR containing complex 1PJ2/ reference 1GZ3), some with a modified nucleotide base (IMP, 3ZCT/8GPB), and others are inhibitors (2T6, 3KC1/4GWZ) including ribose-phosphate complexes (1GPY/8GPB). More rarely, classical bioisosteres such as phosphonates (e.g. AP2, 2VQD/1DV2) or BeF_3^- (e.g. ONP, 1D0Y/1FMW) ligands are found. Cyclic monophosphate replacements can also be found (CMP, 2PW3/1ROR).

In the SO_2 category, the larger fragment $\text{SO}_2\text{-NH-CO}$ is very common, for example (SSA, 1SET/1SES). When mimics occur, hydrogen bonds can be accepted by the oxygen atoms but also donated by the NH, a reminder that neutral hydroxyl of phosphate act also as hydrogen bond acceptors. The carbonyl is usually not seen mimicking the phosphate, but there are exceptions (FG6, 3IVX/2A7X). Other replacements for example sulfonamide, are also found (BII, 2JKM/1MP8; 2DD 2BYI/1BYQ). Sulfonate (SO_3) can like phosphate accept hydrogen bonds (PIN, 1ECF/1ECJ) and also act as a negatively charged replacement. Cyclic sulfonamides are found as well (740, 2R64/4I3Z).

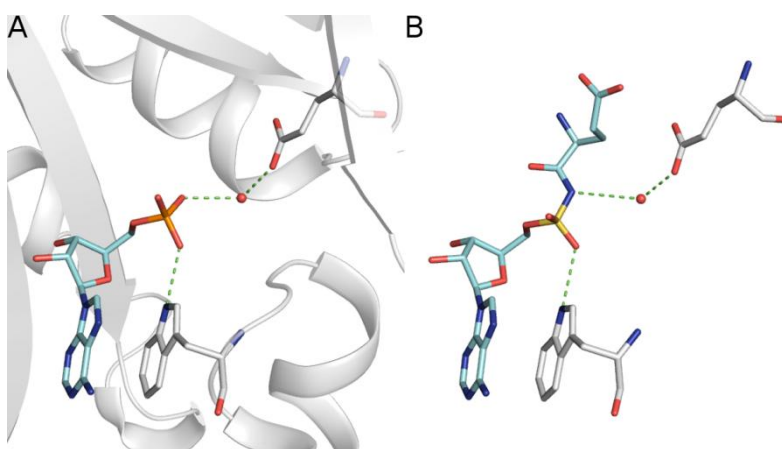


Figure 27 A) 3R96, **Q47510** (*Q47510_ecolx*), 1.85; B) 3R9F, **Q47510** (*Q47510_ecolx*), 2.00; Water is represented by a red sphere.

In the COO group, compounds are very polar, many containing carboxylic acids, and to a lesser extent esters and hydroxyls. One example of different group is dihydrofuran-2(3*H*)-one (A51, 2QFO/1BYQ). More surprisingly, otherwise quite hydrophobic example can be found (0MA, 2QN2/1FA9). Trimethylglycine, a crystallization agent, has a quaternary amine that goes “on top” of phosphate (BET, 4H5W, 1ATR – BET) and is very promiscuous to the dataset. Another basic compound found is another crystallization agent,

EDTA, (EDTA, 2AXN/3QPU), whose positive charge is intramolecularly neutralized. The other group C+O contains compounds with only C and O atoms inside their LSR. Compounds are more diverse than in COO group but appeared to contain less interesting examples.

The CON dataset contains peptides/carbamoyls, discussed below, which seems to be good mimetic of phosphate. A prevalent example is given by GDM, a cyclic HSP90 inhibitor (59c, 4ASA, 1AM1). Often only the carbonyl oxygen is mimicking the interaction of phosphate, for example of the ATP α -phosphate oxygen (LI8, 1Z5M/1H1W; Z48, 3I4B/1J1C). In other examples a water molecule is involved in the interaction (ZRK, 3ZRK/1J1C). Less conventional cases, also found in the CON group, are a tertiary amine (methylnpiperazine) on top of Pi2 (D42, 2B52/1B39), and a nitro group (XKL, 4BQJ/1BYQ).

The cycle group represents many interesting possibilities for medicinal chemists. Among the cycles, cyclohexane 4SP (2C6O/1B38), imidazole 61K (4FKV/1B38), trichlorobenzene (D05, 2B54/1B38), piperazine (D42, 2B52, 1B38), cyclohexanamine (DT4, 2C6L/1B38), triazole (MTW, 2C5Y/1B38) pyrrolidine (2KL, 2WI7/1BYQ) are for example found. More surprising replacements are for example cyanuric acid (RDD, 3BD6/1FA9, an example shown below), ribose (NAI, 1PJ2/1GZ4), morpholine (3Q1, 3QCS/1H1W), benzene (very good example of superimposition, 3RC, 3RCJ/1H1W), nitrophenyl (2XKL, 4BQJ/1BYQ, fluorobenzenesulfonamide (2NQ, 3S2A/1E8X), stacking difluoro (PM1, 1PYE/1B38). Tricyclic ring system containing tertiary amine (2Q9, 4O05/1BYQ), and tetracyclic rings with ether or hydroxyl (KWT, 1E7U/1E8X) are also found.

The remaining part of the dataset contains *tert*-butyl (1YG, 4LYN/1B38) for the C only set; C+N primary amine KJ2 (2BRE/1AM1) or aniline (5DE, 1YDE/1PTW); for example BeS3 (DAA, 1BS1/1A82) or vanadate (VO4, 2QWM/1KAX).

- **Difficulties associated with the categorization**

It is methodologically difficult to define the exact position that corresponds to the local structural replacement. I used spheres to identify the space containing these local

structural replacements. The first issue comes when the individual atoms of the fragments are not well superimposed; the question is then how to define “superimposition” and how much “offset” there can be to still have a replacement. This issue is often occurring for replacements that are located “half-way” between the ribose and the phosphate (see the examples.) In order to address this problem, we used the ShaEP score to eliminate poor structural overlaps, but this leads to the smallest replacements to be eliminated from the dataset (for example the replacement shown in Figure 28 that will appear later is eliminated from the final set because the size of the replacing fragment is too small).

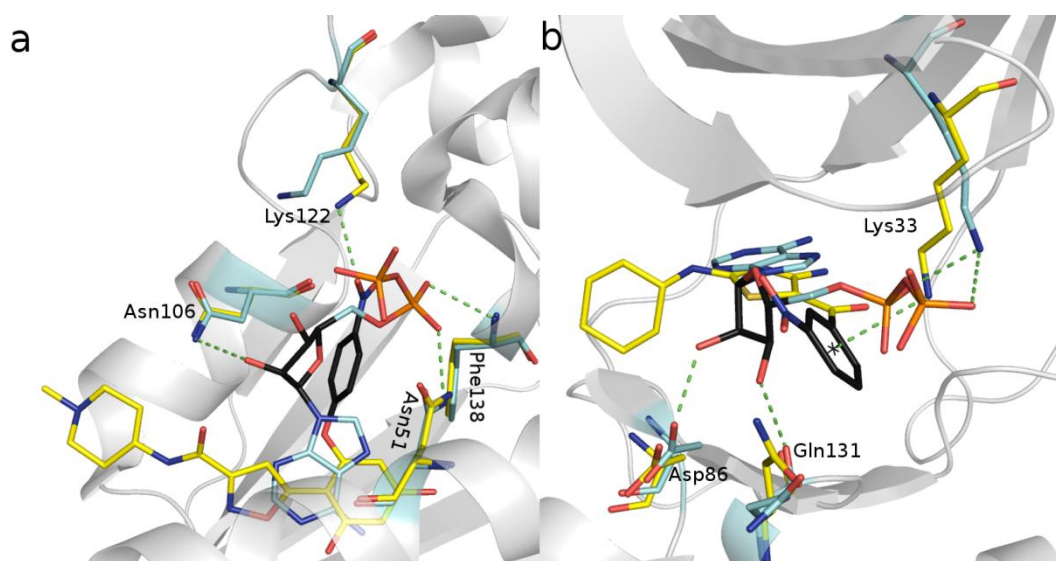


Figure 28 Two examples of a non-clearcut definition of local structural replacements (black atom carbons). Superimposition of the ligand binding site for ADP (cyan)/ nitro-containing ligands (yellow) pairs: **A)** ADP (in 2XK2, 1.95 Å)/ 9UN (in 4B7P, 1.70 Å) bound to heat shock protein HSP90, Uniprot code: **P07900** (HSP90AA1_human) **B)** ADP (in PDB code: 4I3Z, resolution:2.05 Å)/ Z68 (in 3R9N, 1.75 Å) bound to cyclin-dependent kinase 2, **P24941** (CDK2_human) The rest of figures abide by the same format to protein-ligand contact.

In the first example (Fig. 28a) the *para*-nitrophenyl moiety is “between” the ribose and the phosphate of ADP and therefore it is hard to tell which group the nitro group is going to mimic. According to the sphere cutoff criteria, the *para*-nitrophenyl is within the vicinity of both the ribose and the phosphate, but it is a replacement in terms of molecular interactions for neither ribose nor phosphate. The non-replaced interactions are: 2'-oxygen of ADP hydrogen bonding to nitrogen of Asn106 residue; the phosphate O1A and OA2 form hydrogen bonds with main chain nitrogen of Phe138 and side chain nitrogen of Asn151; when the *para*-nitrophenyl compound is bound, it forms hydrogen bond between

the positively charged Lys122 residue and the electron rich nitro group. The phosphate of ADP did not make a charge-reinforced hydrogen bond with adjacent Lys122 residue.

In the second example (Fig. 28b), the *meta*-nitrophenyl moiety is detected by the workflow as “closer” to ribose in terms of distance. However, the *meta*-nitrophenyl moiety is a mimic of the phosphate in terms of interactions, even if this required binding site flexibility (induced-fit): the negative charged phosphates are neutralized by the positively charged Lys33, while Lys33 and the aryl ring of the *meta*-nitrophenyl compound are within cation- π interaction distance (3.9 Å). This illustrates a very difficult case that is not accounted by the method.

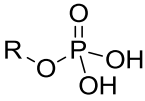
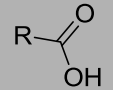
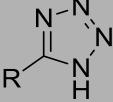
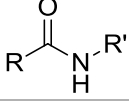
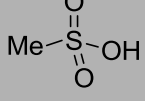
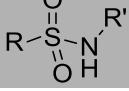
Another comment that can be made on the workflow is that the used SMILES sorting is quite rough. Some fragments could have been put to several categories, but we preferred the option to sequentially sort the SMILES fragments in what we considered the most interesting classes (first taking out the least interesting class, i.e. phosphorus-containing) and to avoid multiple occurrences. We hope the dataset is in this way more manageable. Another issue is the identification of chemical ring structures, which is done by picking SMILES fragments with the letter “1”. The sphere cutoff described previously may lead to a “truncation” of the molecules; in particular rings that are offset are not detected as such in the SMILES codes. We are aware of the problem but solving it is quite complex.

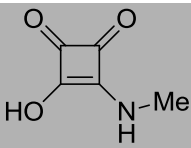
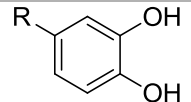
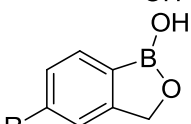
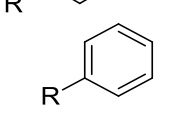
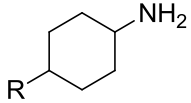
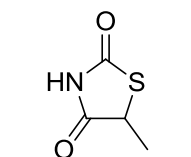
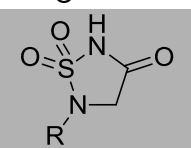
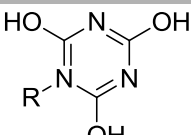
The last concern is a potential bias in the dataset: firstly, e.g. kinases are over-represented, which prevents us from making meaningful statistics of the binding site amino acids. This issue would require proper sorting of the proteins’ folds. Preliminary attempts with sequence-based fold recognition did not succeed. A way to get ahead may be to extract the SCOP/Pfam codes from the PDB files for example to sort proteins according to families. Secondly, the frequencies of the type of replacements found represent the work of the crystallographers and the directions privileged by chemists; it is therefore not surprising those well-known replacements such as phosphate and ester/carboxylate look like to be the most common.

5.2 Previously unrecognized structural replacements of phosphate

As mentioned in the Review of Literature, phosphate groups are highly polar and singly or doubly negatively charged groups. In our sample of phosphate groups bound found the PDB, a high number of coordinated metals were found (see Fig. 22 and Table 5). Previously identified isosteres of phosphate (presented in the Introduction) that conserve charge were found to be the most prevalent, without any surprise: for example carboxylate or phosphorus-containing replacements (numerical data about occurrence of the relevant SMILES code in the dataset is provided in Fig. 25). There are also quite many replacements as sulfonyl- or ester-containing compounds. Tetrazole replacements are not found in the dataset, but very few tetrazoles are found in the PDB altogether. A summary of the pK_a values of the common phosphate replacements is presented in Table 6 in grey background along with their hydrogen bonding properties (neutral form).

Table 6. *Examples of the pK_a and hydrogen bonds from data set. Hydrogen bond donating (HBD), accepting (HBA), or groups capable of both (HBA/D) are counted for the fragments in their neutral forms. Previously known common phosphate isosteres are shown with grey background. Fragments are ordered by growing size.*

Chemical group		pK_a	HBA/D	only HBA	only HBD
	AMP ADP ATP	1.54 ^a 6.31 ^a 6.5/7.1 ^a			
		0.5-4.76 ^b	1	1	0
		4.90 ⁱ	0	3	1
		>15 ^h	0	1	1
		-1.90 ^c	1	2	0
		10.7 ^d	0	2	1

	2.3 ^e	1	2	1
	9.48 ^f	2	0	0
	7.4 ^g	1	1	0
		0	0	0
	10.64 ^k	0	0	1
	5.0-6.5 ^j	0	2	1
	15 ^l	0	3	1
	6.88 ^m 11.40 ^m 13.5 ^m	3	0	0

a (Elliott et al., 2012); **b** http://en.wikipedia.org/wiki/Carboxylic_acid; **c** http://en.wikipedia.org/wiki/Sulfonic_acid; **d** (Şanlı, Şanlı, Özkan, & Denizli, 2010); **e** (Elliott et al., 2012); **f** (Ganellin, 1977); **g** (Ellis, Palte, & Raines, 2012); **h** <http://en.wikipedia.org/wiki/Amide>; **i** (Satchell & Smith, 2002); **j** http://www.cambridgemedchemconsulting.com/resources/bioisoteres/acid_bioisoteres.html; **k** (Hall, 1957); **l** (Sparks et al., 2007); **m** http://en.wikipedia.org/wiki/Cyanuric_acid.

Table 6 contains also selected examples of the local structural replacements that we found more surprising, and that we name here “previously unrecognized” structural replacements. These “previously unrecognized” structural replacements could be divided into four categories: (1) acidic groups of elevated pK_a that in the context of a binding site are likely to be ionized (no quantum chemistry calculations are performed in this work, only evaluation based on the PDB coordinate); (2) polar non-acidic groups with charge delocalization through conjugated bonds or inductive effects via σ -bonds; (3) aromatic non-polar groups; (4) basic groups. The two last groups represent intriguing possibilities to medicinal chemists. How such structural replacements are managed by the binding site is illustrated on selected examples below, as well as in Figures 29-32. Figures 29-32

illustrate the local structural replacements of phosphate by carbamoyl, carboxycatechol, pyridine, and primary amine moieties. The molecular mechanisms of molecular recognition discussed in this thesis are summarized in Section V.4.

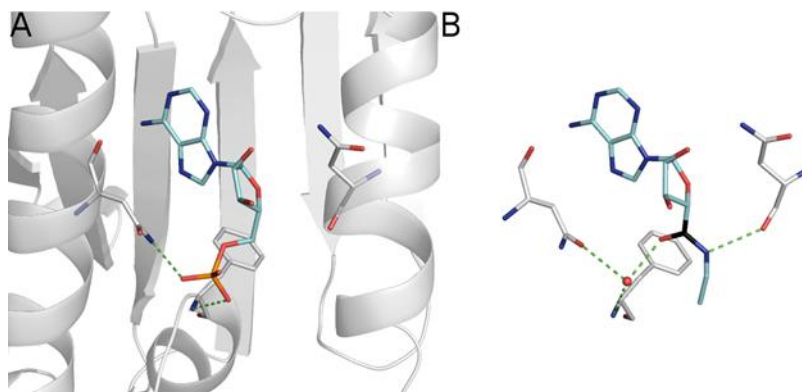


Figure 29 *A) 1TBW, P41148 (ENPL_canfa), 2.19 Å; B) 1QY5, P41148 (ENPL_canfa), 1.75 Å. Water is represented by a red sphere. All the following figures are labeled as follows: the protein is shown as a white ribbon, the concerned amino acids and ligands are shown as sticks, the ligands are colored as cyan and the corresponding replacement moieties are shown as black. This example was excluded by the empirically set ShaEP filter used to compile the final data.*

In the dog endoplasmic protein (alternatively named as heat shock protein 90 kDa beta member 1, HSP90; Figure 29A), (Soldano et al., 2003), (Immormino et al., 2004), (Dollins et al., 2007) two phosphate group distal oxygens of AMP are hydrogen-bonded to the amide nitrogen of Asn107 (distance 3.03 Å) and to the main-chain nitrogen of Phe199 (3.09 Å). In Figure 29B, the phosphate of the AMP analog *N*-ethyl-5'-carboxamidoadenosine has been replaced by a carbamoyl containing fragment. In this structure, the carbonyl group of the replacement is not in direct contact with the amide group of Asn107, but rather interacts through an intercalated water molecule, which itself possibly interacts with the (flipped) amide group of Asn107. In addition, the nitrogen of the replacement is within hydrogen bonding distance (3.02 Å) from the main-chain oxygen of Asn162.

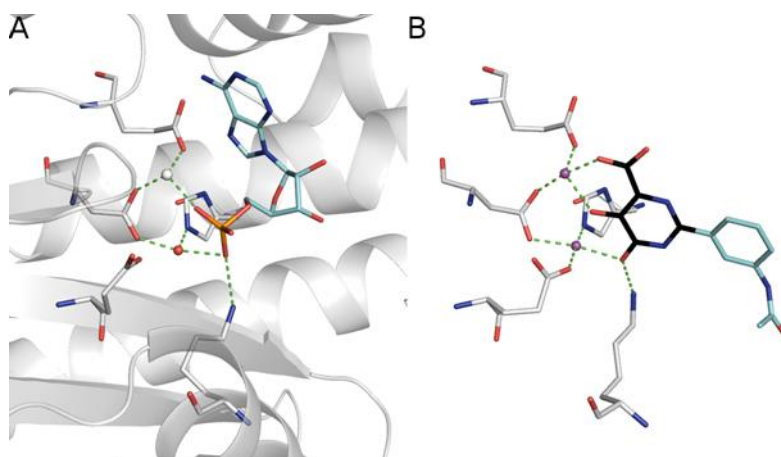


Figure 30 *A) 3HW5, Q9Q0U9 (PA_I96A0), 1.81 Å; B) 4E5J, Q5EP34 (Q5EP34_9INFA), 2.35 Å; magnesium and manganese ions are represented by white and magenta spheres, respectively.*

The binding site of the influenza A polymerase bound to AMP is highly charged, with three negatively and two positively charged amino acids (Figure 30A). The interaction of the AMP phosphate is mediated by a magnesium ion as well as a salt bridge to Lys134. One water molecule is also involved in this interaction network (Zhao et al., 2009). In Figure 30B, carboxycatechol acts as a replacement of the phosphate, while the rest of the ligand is different. Two Mn^{2+} ions are found in the X-ray structure; one is coordinated by both of the hydroxy groups and the other by one of the hydroxy groups and by the carboxyl group. The carboxycatechol moiety likely carries two negative charges since chelation by catechol usually happens in their ionic forms (electron withdrawing groups enhance ionization of catechols) (Xhaard et al., 2008), (Hikaru Harada, 1971). Other residues coordinating the Mg^{2+} are His41, Asp108, and Glu119, as well as Glu80 and Asp108 (DuBois et al., 2012).

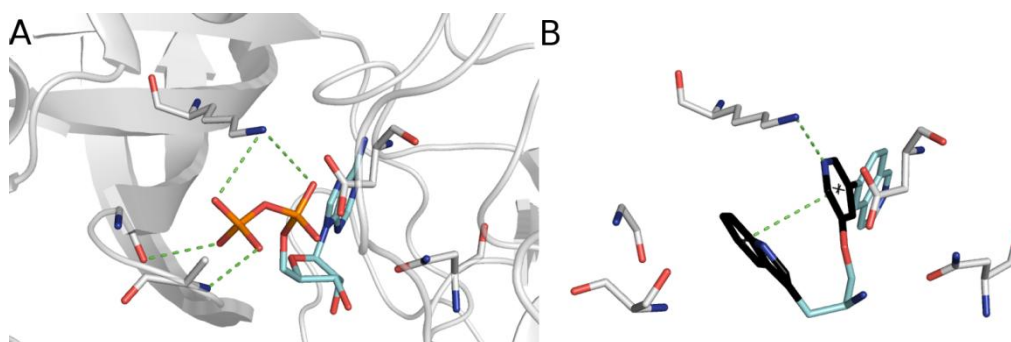


Figure 31 *A) 1JBP, P05132 (KAPCA_mouse), 2.20 Å; B) 2F7E, P00517 (KAPCA_bovin), 2.00 Å.*

ADP bound to the mouse cAMP-dependent kinase is shown in Figure 31A (Madhusudan et al., 1994). The replacement is exemplified by the structure of a bovine ortholog bound to a bi-aromatic ligand that contains indole and pyrimidine rings (Figure 31B). The two rings are stacked (stacking distance 4.6 Å) in the replacement (Li et al., 2006). The proximal phosphate is replaced by the 3,5-substituted pyridine, whose nitrogen atom forms a hydrogen bond to Lys72 (2.8 Å). The indole ring is intramolecularly stacked on the pyridine ring as well as the Lys72 side chain with hydrophobic interaction. In both cases, Lys72 is involved in a charge-reinforced ion paired to Glu91 (2.6 Å). This type of replacement is very interesting because there are many occurrences of intramolecular stacking of compounds in the dataset, a fraction of which occur at the phosphate binding site. In the case presented, adding a ring during study of SARs the preserve the biological activity but do not lead to more potent compound.

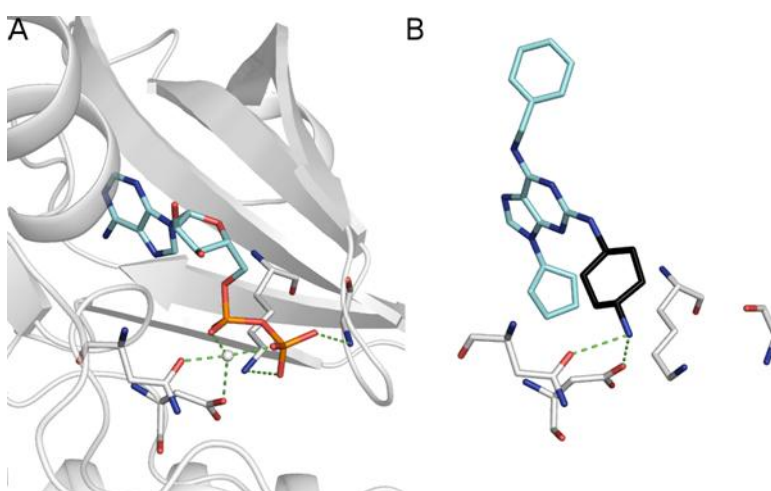


Figure 32 *A) 3QHW, P24941 (CDK2_human), 2.17 Å; B) 1G5S, P24941 (CDK2_human), 2.61 Å. The magnesium ion is represented by a white sphere.*

The last example is cyclin-dependent kinase 2 (Bao, et al., 2011) bound to ADP (Figure 32A). A magnesium ion is hexacoordinated by two of the oxygen atoms of the phosphates (O2A and O1B) as well as by side chain oxygens of the Asn132 and Asp200 residues. The primary amino group of Lys33 interacts with both phosphates' oxygens but it is closer to the O1A oxygen. In Figure 32B, the phosphate replacement is a cyclohexylamine moiety whose primary amino group sits exactly at position of the magnesium ion, forming a charge-reinforced hydrogen-bond with Asp200. Lys85 is pushed away in the replacement.

5.3 Structural replacement of ribose

Of important note, ribose in the context of this study means a ribose linker as found in AMP, ADP and ATP, i.e. connected at the 1' and 5' positions. The majority of the ribose linkers found in this study is in the β -pyranose form.

A ribose linker is uncharged under normal conditions, and although it is very polar, it contains an aliphatic “edge” that can be accommodated in a binding site. This versatility seems to be reflected in the variety of groups that we found were able to work as linkers (Figure 33). To our best knowledge, ribose structural replacements have not been characterized so far, possibly because replacing ribose is less problematic to organic chemists than replacing phosphate. Quite often, the ribose is exposed to solvent when a nucleotide binding site is a crevice in the protein. Seemingly, the same types of replacements could be found for ribose and phosphate.

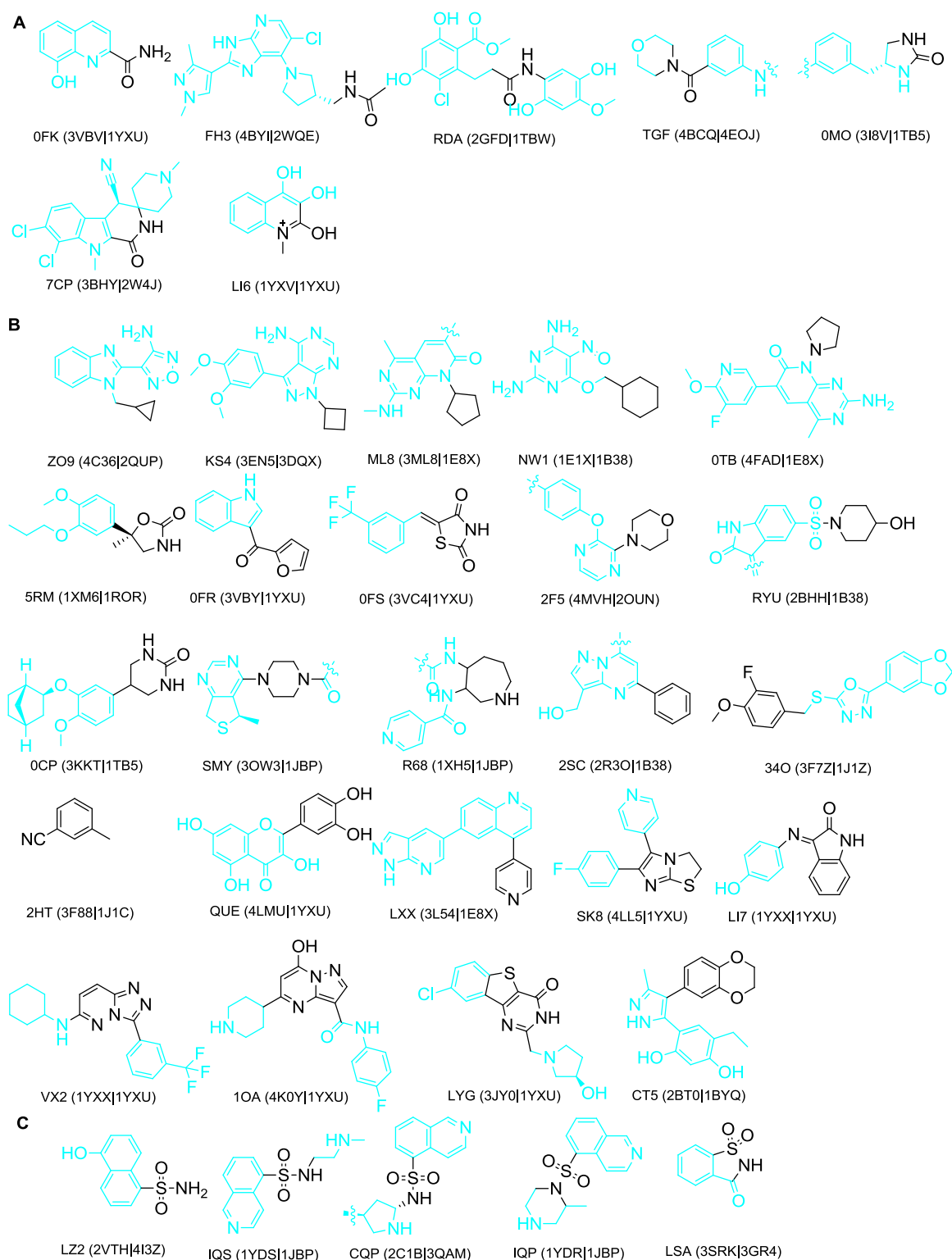


Figure 33. *Local structural replacements of ribose (black) (A) CON (B) cycle (C) SO₂. All example were labeled as following ways: ligand three letter or digital code (code of protein who complex with the ligand| code of reference protein).*

Three examples of ribose replacements are presented in more detail below: a phenol, a series of alicyclics, and an amide.

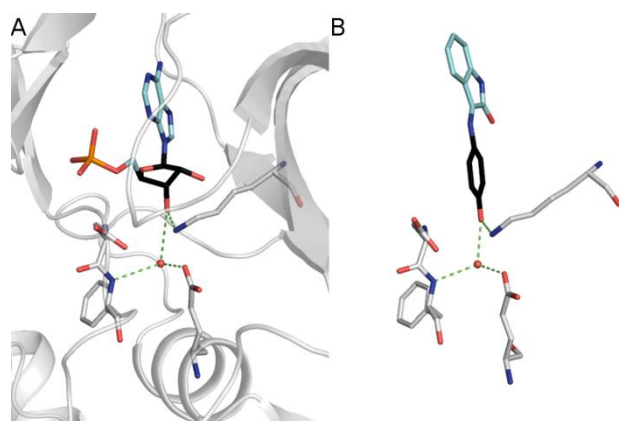


Figure 34 A) 1YXU, P11309 (*pim1_human*), 2.24 Å; B) 1YXX, P11309 (*pim1_human*), 2.00 Å

Figure 34 shows the proto-oncogene kinase Pim1 (Kumar et al., 2005). In Figure 34A, one water molecule interconnects the AMP ribose 3' oxygen. This water molecule interacts with the Phe187 backbone nitrogen, the Glu89 side-chain, and the Lys67 side-chain. Figure 34B shows a very conservative replacement of a ribose by a phenol, which is mediated by a water molecule. The phenol oxygen takes exactly the same place of the 3'-oxygen of ribose.

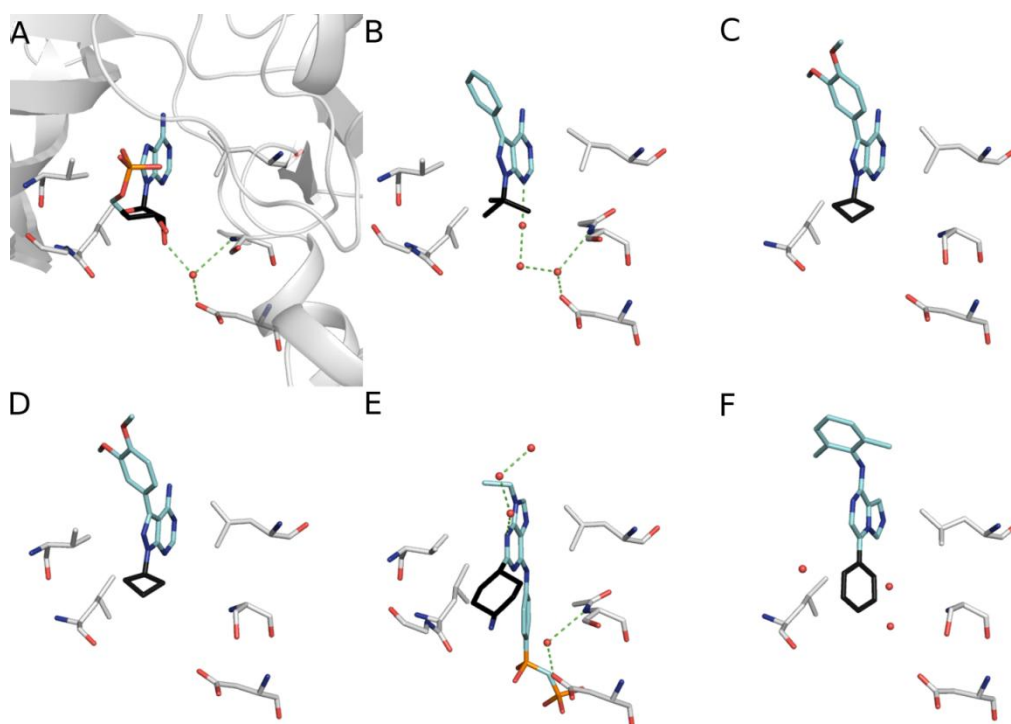


Figure 35 A) 3DQX, P00523, (SRC_chick), 2.30 Å; B) 1QPE, P06239, (LCK_human), 2.00 Å; C) 3EN5, P00523 (SRC_chick), 2.66 Å, D) 3EN4, 2.55 Å, P00523, SRC_chick, E) 2BDF, 2.10 Å, P12931 (SRC_human), F) 2ZYB, 2.55 Å, P06239, (LCK_human).

Figure 35 shows three types of kinases sharing the same fold, two from human src (Dalgarno et al., 2006) and lck (Zhu et al., 1999), (Ozawa et al., 2008) and one being chicken c-Src kinase (Azam, Seeliger, Gray, Kuriyan, & Daley, 2008). In Figure 35A, the 2'-oxygen of ribose is hydrogen bonded to the side chain oxygen of Asp348 via an intercalated water molecule. This water molecule is also hydrogen-bonded to the main chain nitrogen of Ser345. Figures 35B-E present examples where the ribose is replaced by hydrophobic groups, such as *tert*-butyl (panel B), cyclobutyl (panel C), cyclopentyl (panel D), cyclohexyl (panel E) and phenyl (panel F). In panel B and E this water molecule is also observed. In panel C and D, water molecules are not found. The reason is probably the lower resolution of 2.66 Å in panel C instead of 2.00 Å in panel B. In F, the conserved water is not visible but more bulk water is seen. The binding site seems unchanged but solvent reorganization seems to accompany the replacements although the somewhat low resolution of some examples may prevent a complete understanding.

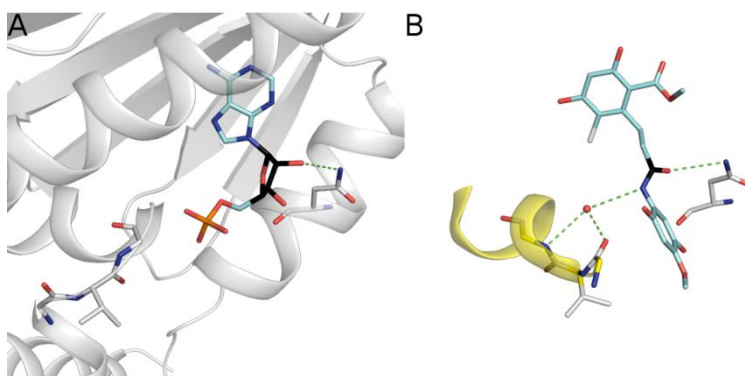


Figure 36 *A)* 1TBW, P41148, (ENPL_canfa), 2.15 Å; *B)* 2GFD, P41148, (ENPL_canfa), 2.30 Å

Figure 36 shows the dog glucose-regulated chaperone protein (Immormino et al., 2004). In panel A, the 2'-oxygen of the ribose of AMP hydrogen bonds with residue of Asn162 (3.52 Å) and there is no interaction with Gly196. In panel B, a different ligand is bound and an amide replacement of the ribose is shown. The amide nitrogen forms a hydrogen bond to the Gly196 main chain carbonyl through a water bridge (3.26 Å with Gly196 nitrogen, 2.82 Å with Gly198 oxygen) (Immormino et al., 2009). This interaction “pulls” Gly196 closer to the binding site that likely induce a loop to helix (panel A to B) conformational change about the three amino acids (Gly196, Val197 and Gly198).

5.4 Molecular mechanisms involved in the recognition of local structural replacements

To study the flexibility, the reference binding sites in the reference proteins were defined by any amino acid with at least one atom within 4.5 Å of the bound AMP, ADP, ATP or POP. Equivalent positions were retrieved based on the sequence numbering in the LSR-containing proteins, and only when an exact match of all amino acid types occurred flexibility could be studied (i.e., only a fraction of the dataset is included here). The result shows that while even large loop movements may occur; binding sites typically do not undergo generalized changes. (Figure 37)

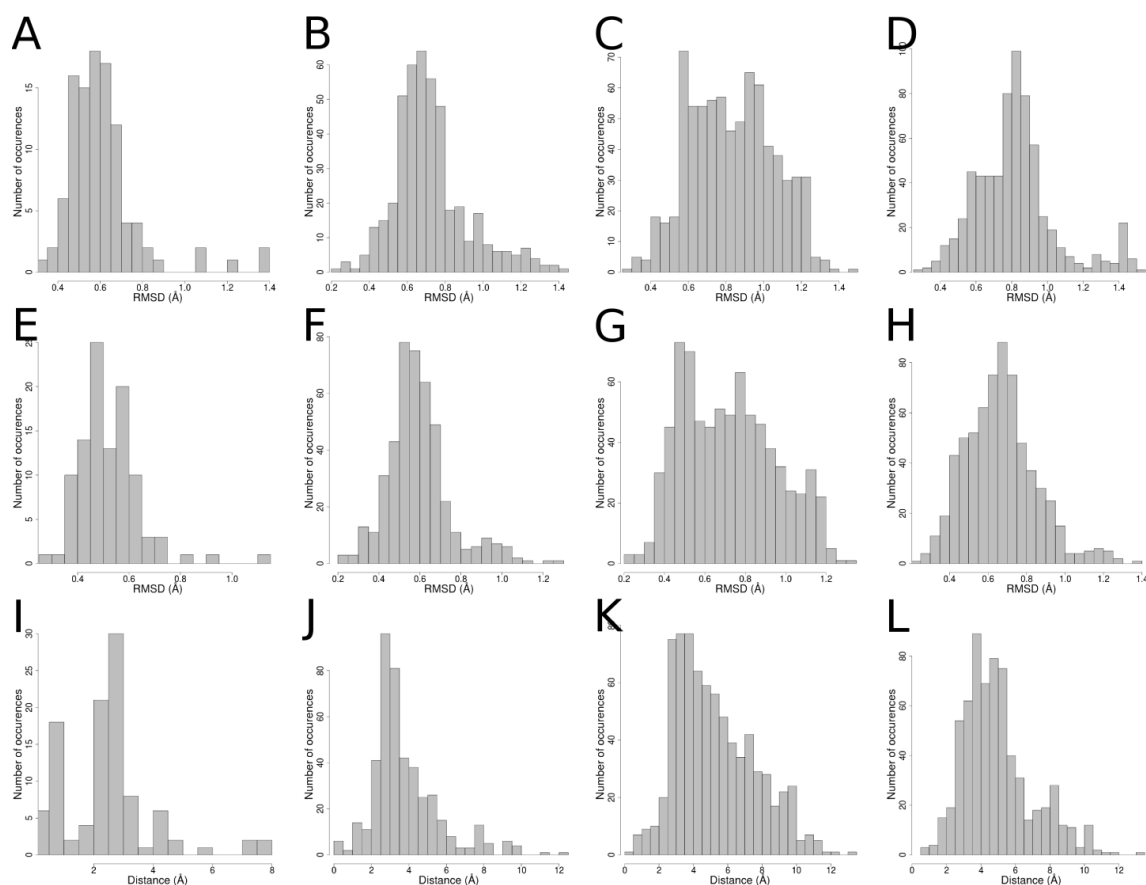


Figure 37. Distribution of the binding sites flexibility for (A-D), all atom RMSD (Å); (E-H), Ca-carbons RMSD (Å); (I-L) longest displacement (Å). RMSD are for the datasets (A, E, I) POP; (B, F, J) AMP; (C, G, K) ADP and (D, H, L) ATP datasets.

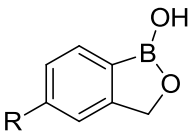
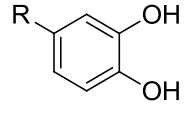
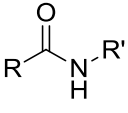
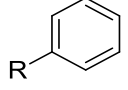
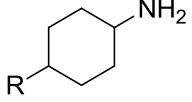
From the examples presented in the previous section, it can be concluded that it is the molecular interactions between ligands and binding site residues that anchor the ligand. The protein binding sites accommodating ligands are “induced-fit” processes where the residues reorganize themselves to host ligands in a favorable way. Different ligands will drive residues in different directions; some of them may lead to a minor change of the position while others will dramatically be altered. Considering the examples given previously in Figures 29-36, as well as our analysis, it is very seldom that the ideal case of high conservatively mimic of interactions is seen. The molecular mechanisms that follow structural replacements are multiple and may involve:

- Main chain and side-chain protein conformational change (Figure 36)
- New interactions mediated by water molecules (Figures 34 and 35)
- New interactions mediated by metals/ions (Figure 30)
- Replacement of the ligand by a metal
 - *vice versa*, replacement of a metal by the ligand (Figure 32)

- Replacement of the ligand by a bridging water (Figure 29)
 - *vice versa*, replacement of a water by the ligand
- Replacement of the ligand by a fraction of the binding site (Figure 28b)
- Rearrangement of “bulk” solvent to accommodate hydrophobic groups (Figure 35)
- Ligand collapse to accommodate hydrophobic groups; intramolecular hydrophobic stacking (Figure 31)

For the phosphate groups, local structural replacements seem to be associated with preferred mechanisms. The combinations identified in this study are summarized in Table 7.

Table 7. *The exemplified structural replacements with preferred mechanisms.*

Structural replacement	Structure	Mechanism in the example given	Figure	Properties
benzoxaborole		direct metal chelation by the boron atom (likely ionic form)	-	acidic with elevated pK_a
catechol		direct metal chelation (likely ionic form)	30	
amide		direct replacement (mediated by a water bridge in example)	29, 36	polar non-acidic groups but with charge delocalization through conjugated bonds or inductive effects
		hydrophobic ligand collapse	31	apolar aromatic/aliphatic
		solvent-facing exchange		
		the base is “replacing” the metal coordinated by the nucleotide	32	basic groups

5.5 Discovery of compounds binding to MDO1

The second part of the thesis deals with the design and synthesis of potentially bioisosteric replacements of phosphate-distal ribose of ADPr. This includes ligand based design, synthesis, biological activity experiments, and retrospective docking analysis. The biological activity experiments were done in Oulu (group of Dr. Lari Lehtiö) using differential scanning calorimetry.

5.5.1 Modeling MDO1 – structural analysis

When this project started, only the *apo* form of the human MDO1 was available (PDB code: 2X47) but it was not suited for compound design. However, the ortholog structure of *Archaeoglobus fulgidus* bound to ADP-ribose has been reported (PDB code 2BFQ) (Karras et al., 2005). A homology model of the bound conformation of human MDO1 was constructed based on this template using the software Swissmodeller (Biasini et al. 2014). It was denoted as the *holo* conformation. The model was made such that the crystallized ADPr conformation, borrowed from the template, complementarily fit the MDO1 binding site (Figure 38B). A pocket is formed by helices 8, 11 and 13, as well as lined by beta strands 2, 7 and 8 (Figure 38B). The pocket can be divided into five regions depending on the fragments of the bound ADPr that they accommodate (Figure 38C): S1, the adenine; S2, the proximal ribose; S3, the proximal phosphate; S4, the distal phosphate; and S5, the distal ribose. The macrodomain pocket is L-shaped and the turn occurs in the S2 region.

The major differences between the *holo* and the *apo* MDO1 conformations are localized in two regions (Figure 38A). Firstly, the S5 sub-pocket is more open in the *apo* conformation. The most flexible loop slice from Gly270 to Phe272 remarkably moved apart from the binding site compared to the *apo* conformation, leading to 4.21 Å variations (RMSD) of the main chain. In the *holo* conformation, the residue of Phe272 flipped down towards the binding site and the whole loop also forms a compact space to accommodate ligands. Secondly, in the sub-pocket 1, the Phe306 residue (yellow labeled) of the *apo* conformation intersects with the adenosine ring, while it is parallel in the *holo* conformation. In addition, the isopropyl group of the Val183 residue (yellow labeled) clashes with the adenosine ring. Hence, in the *apo* conformation these two residues

arranged in a way that the S1 sub-pocket was closed and hence it is not able to bind to the adenosine.

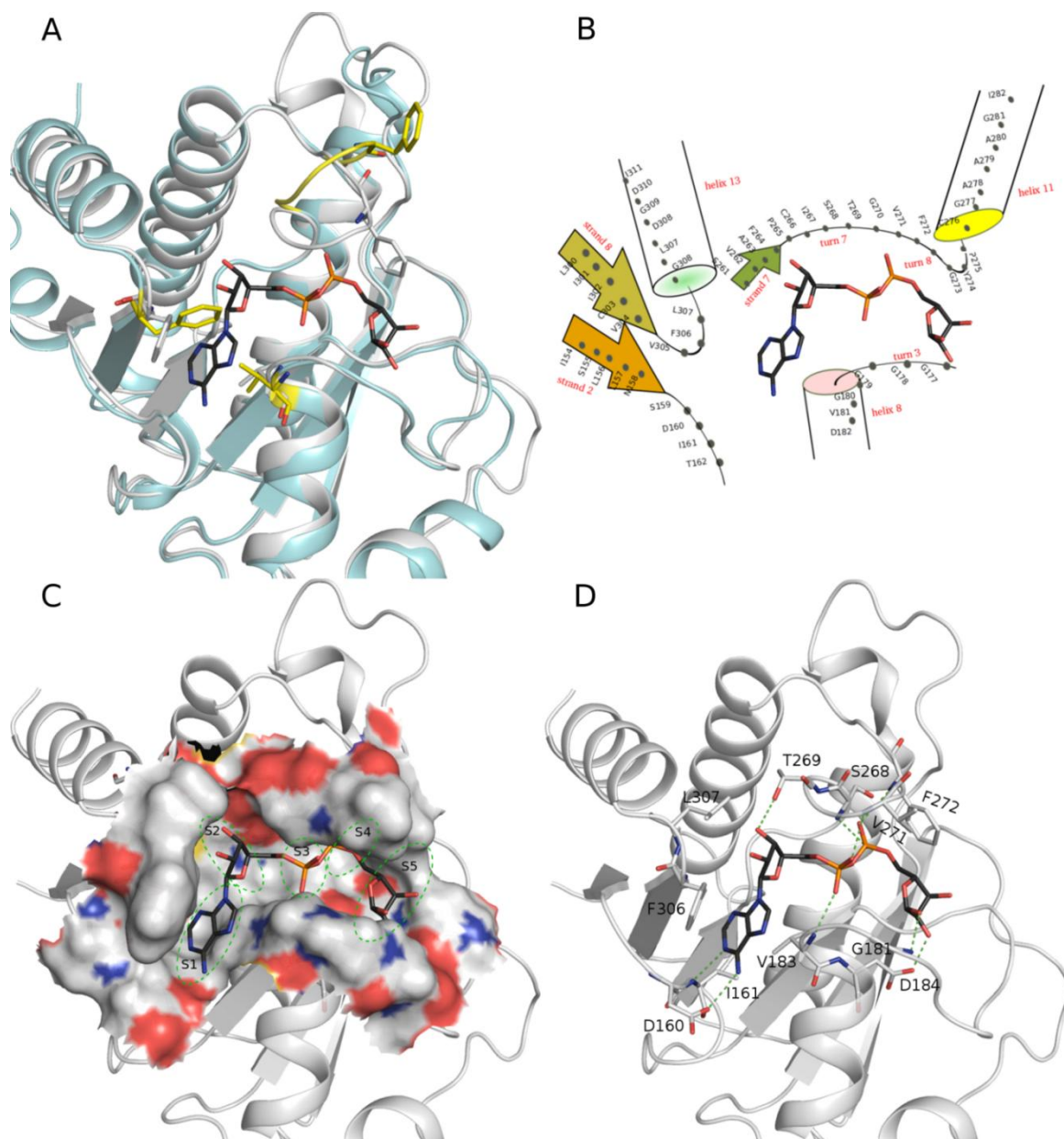


Figure 38 *Three dimensional model of human macrodomain protein 1 (MDO1) complexed with ADPr. (A) Structural superimposition of the apo (cyan, PDB code: 2X47, Val183, Phe306. Phe272, yellow) and holo (homology model, white) states. (B) Schematic presentation of the holo ADPr binding site. (C) The surface of the holo state's binding pocket. Green dash eclipses label the individual subpockets from S1 to S5. (D) Detailed stereo view of the interactions at the ADPr binding site of the holo state. Green dash lines represent the hydrogen bonding.*

The interactions of ADPr with MDO1 is depicted by Figure 38C-D. The S1 region

accommodates the adenine ring. The purine ring is sandwiched between the planar phenyl group of Phe306, interacting with π - π interaction, and the aliphatic side-chain of Val183. The NH₂ group of adenine (N6) is donating a hydrogen bond to the side chain carboxyl oxygen of Asp160. In addition, the N1 nitrogen from the Watson-Crick edge of adenine accepts a hydrogen bond from the main-chain nitrogen of Ile161. The S2 region holds the proximal ribose. The proximal ribose is surrounded by Pro265, Asp310, Cys266 and Thr269. O3' of the ribose donates a hydrogen bond to the side-chain of Thr269. The S3 region accommodates the proximal phosphate, which has a first pK_a of 1.54 and therefore is likely to be negatively charged. The negative charge of the phosphate is neutralized by the positively charged macrodipole at the N-terminus of helix 8. This macrodipole is generated by the sum of the dipolar contributions from the helical backbone. In addition, on the top of this helix, the NH groups of Val183 and Gly182 main-chain donate the hydrogen bonds to oxygen of the phosphate. In the S4 region, the distal phosphate is trapped by a part of the turn, which links beta strand 7. Specific hydrogen bonds are also formed between the amide nitrogen amino acid Ser268, Gly270 and Val271. The distal ribose moiety attached to the distal phosphate group is located in the S5 region, which is formed, by Asn174, Asp184, Gly181 and Phe272. The O1' and O2' of ribose donate hydrogen bonds to the main-chain nitrogen of Gly181 as well as to the side-chain of Asn174 and Asp184 respectively. In addition, the side-chain of Phe272 is close to the ribose.

5.5.2 Synthesis of ADP analogs

To investigate how phosphate isosteres are accommodated at the MDO1 site experimentally, we designed a set of ADP analogs. Among these analogs, we kept the adenosine and ribose moieties unchanged while the phosphates were replaced by potential bioisosteres, linking variety of substituents. The phosphate replacer can be regarded as a linker. Two series of analogs were designed: the squaryldiamide (Scheme 1) and the amide (Scheme 2 and 3) series. Nine compounds are based on a squaramide linker and six compounds based on an amide linker. At the beginning of the study, besides these two series, I also planned a series of compounds with sulfonamides as mimics of phosphate, (Gajewiak & Prestwich, 2006), (Dong et al., 2010) but these were not prioritized and were not synthesized.

5.5.2.1 Chemistry

- **Scaffold**

We started on the synthesis of the target compounds by treating diethyl squarate (**4**) with various aliphatic, aromatic and benzylic amines in ethanol to give nine differently substituted alkoxy amino squarates (**5-13**) in good 71-92% yield (X. Zhang et al., 2013). Only mono-substituted products were formed in this reaction. (Niewiadomski et al., 2010) In a similar fashion, the acetonide-protected 5'-amino-5'-deoxyadenosine (**3**) was reacted with the alkoxy amino squarates **5-13**. This transformation produced the acetonide-protected and unsymmetrically 3,4-substituted cyclobut-3-ene-1,2-diones **14-22** that were subsequently deprotected in the presence of aqueous trifluoroacetic acid (TFA) to give a set of nine different 3-(5'-amino-5'-deoxyadenosine)-4-(alkylamino)cyclobut-3-ene-1,2-diones (**23-31**) in 83-90% yields.

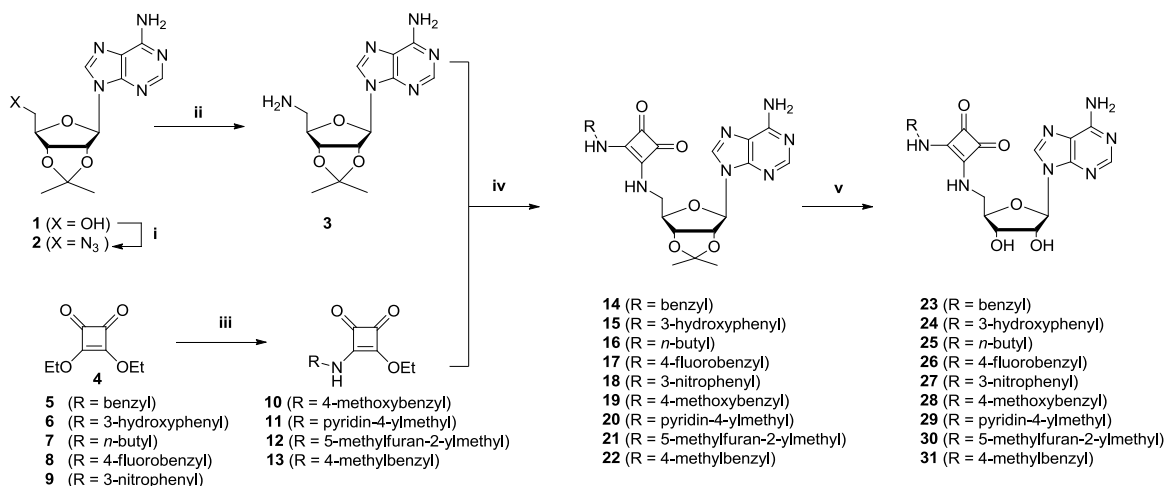
- **Squaryldiamide series.**

Squaric acids and squaryldiamides have been reported as bioisosteres of phosphate (Seio, Miyashita, Sato, & Sekine, 2005) though have not been that successful so far (Moreau et al., 2013). I synthesized nine ADP analogs using squaryldiamide as a linker/core. Adenosine is on the one side of the asymmetrical squaryldiamide, and on the other side different substituents were attached, such as *n*-butyl, phenyl, benzyl, pyridin-4-yl-methyl and 5-methylfuran-2-ylmethyl with the aim to exploit optimal interactions. Purification of the final products were problematic with SiO₂ column chromatography but amine functionalized SiO₂ columns (NH-KP, Biotage) allowed us to successfully purify the synthesized compounds.

The compounds of the first series, squaryldiamides **23-31**, were synthesized in three reaction steps starting from diethyl squarate (**4**) and two different amines (Scheme 1). First, the key building block, acetonide-protected 5'-amino-5'-deoxyadenosine (**3**) was synthesized. We treated 2',3'-*O*-isopropylideneadenosine (**1**) first with diphenylphosphoryl azide (DPPA) and 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU), and subsequently with sodium azide and 1,4,7,10,13-pentaoxacyclopentadecane (15-crown-5) in 1,4-dioxane under reflux to yield acetonide-protected 5'-azido-5'-deoxyadenosine (**2**) in 92% yield. Subsequently, the azido intermediate (**2**) was converted to the desired key building block

(**3**) by using a modified Staudinger reaction in the presence of polymer-bound triphenylphosphine. (Ayesa, Samuelsson, & Classon, 2008) Acetonide-protected 5'-amino-5'-deoxyadenosine (**3**) was obtained in 70% yield.

Scheme 1. Synthesis of the squaryldiamide series.^a



^a Reagents and conditions: i) DPPA, DBU, NaN₃, 15-crown-5, 1,4-dioxane, rt to 110 °C, overnight, 92%; ii) triphenylphosphine (polymer-bound), THF, H₂O, rt, 4 h, 70%; iii) amine, EtOH, 0 °C to rt, 2 h, 71-92%; iv) DIPEA, EtOH, rt, 5 h to overnight, 79-91%; v) TFA/H₂O (4:1), rt, 5 h, 83-90%.

• Amide series

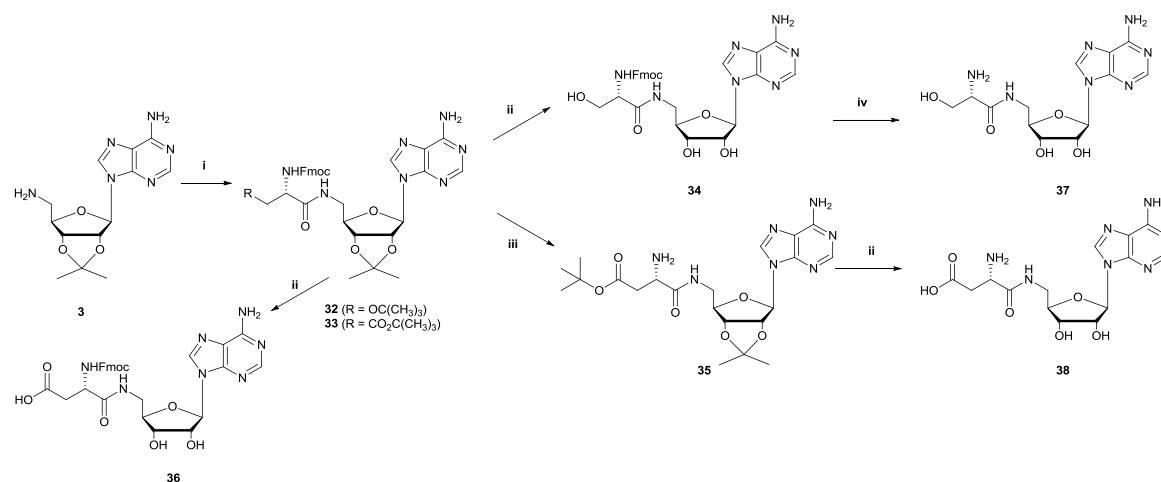
Six amide-linked ADP analogs were synthesized. Compound **37** was aimed at validating the similarity of hydrogen bonding between phosphate and hydroxy, while compound **38** contains a carboxylate group designed to examine the charge similarity to ADPr. The Fmoc-protected compounds **34** and **36** were also tested in order to find a fluorescence marker of that protein to facilitate screening. They are the precursors of **37** and **38**. Tetrazole is a bioisostere of the carboxyl group while the carboxyl group is able to structurally replace phosphate; that was the motivation to synthesize compound **41**, and as a negative control compound **42**.

The synthesis of the amide series was started by coupling amine **3** with *N*-hydroxysuccinimide (OSu)-activated ester and fluorenylmethoxycarbonyl (Fmoc)- and *tert*-butyl protected amino acids (Fmoc-Ser(*t*Bu)-OSu and Fmoc-Asp(O*t*Bu)-OSu, respectively) to give compounds **32** and **33** in 76 and 75% yields, respectively (Scheme 2). Deprotection of the acetonide- and the *tert*-butyl groups of compounds **32** and **33** was performed by a mixture of TFA/H₂O/CH₂Cl₂ (5:1:4) to give the Fmoc-protected amino acid derivatives **34** and **36**, both in 90% yield. The serine derivative **37** was obtained in 43%

yield after deprotection of the Fmoc group with diethyl amine (DEA) in *N,N*-dimethylformamide (DMF). The aspartic acid analog **38** was prepared with a different deprotection strategy. First, the Fmoc group of **33** was deprotected with DEA in CH_2Cl_2 giving the amine **35** in 59% yield. Subsequently, the rest of the protecting groups of compound **35** were removed with TFA/ H_2O / CH_2Cl_2 (5:1:4) to give the product **38** in 81% yield.

The peptide coupling reactions to prepare compounds **37** and **38** were conveniently performed with OSu-activated amino acids. Commercially available OSu-activated starting materials were not available for compounds **41** and **42**. Therefore, HBTU was used as a coupling reagent of the starting material carboxylic acids in these reactions.

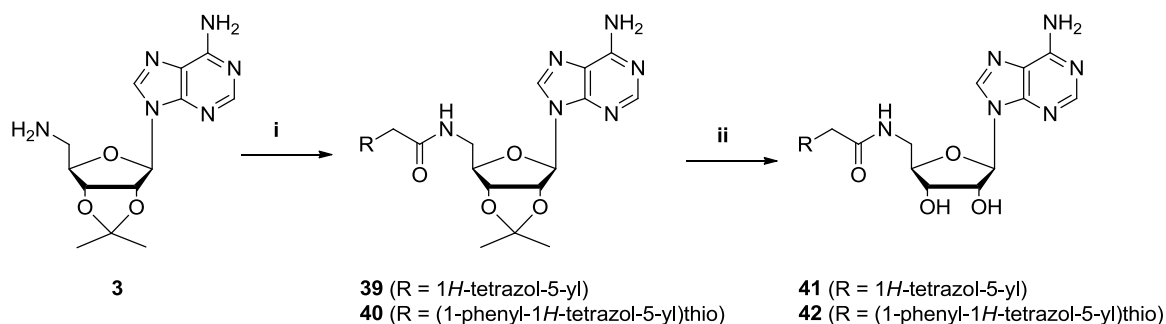
Scheme 2. Synthesis of the amino acid series.^a



^a Reagents and conditions: i) Fmoc-Ser(*t*Bu)-OSu or Fmoc-Asp(*Ot*Bu)-OSu, MeOH, rt, 2 h, 75-76%; ii) TFA/ H_2O / CH_2Cl_2 (5:1:4), rt, 5 h, 59-90%; iii) DEA, CH_2Cl_2 , rt, 2 h, 59% yield; iv) DEA, DMF, rt, 2 h, 43%.

In the syntheses of the two tetrazole analogs **41** and **42**, amine **3** was first reacted with either 1*H*-tetrazole-5-acetic acid or [(1-phenyl-1*H*-tetrazol-5-yl)thio]acetic acid using *N,N,N',N'*-tetramethyl-*O*-(1*H*-benzotriazol-1-yl)uranium hexafluorophosphate (HBTU) as a coupling reagent to yield the amides **39** and **40** in 80 and 85% yields, respectively (Scheme 3). Subsequently, the acetonide-protecting groups of amides **39** and **40** were removed with TFA/ H_2O / CH_2Cl_2 (5:1:4) to give the products **41** and **42** in 83 and 77% yields, respectively.

Scheme 3. Synthesis of the tetrazoles.^a



^a Reagents and conditions: (i) 1*H*-tetrazole-5-acetic acid or [(1-phenyl-1*H*-tetrazol-5-yl)thio]acetic acid, HBTU, DIPEA, DMF, rt, 4 h, 80-85%; (ii) TFA/H₂O/CH₂Cl₂ (5:1:4), rt, 5 h, 77-83%.

• Purification

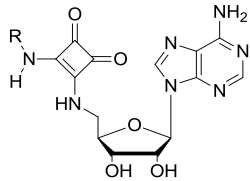
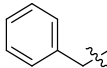
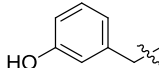
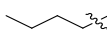
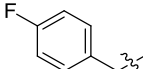
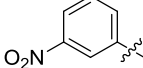
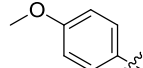
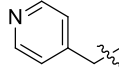
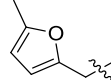
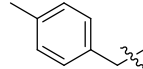
For most of the compounds, the final step of purification was always a challenge since the compounds are really hydrophilic. When the reaction mixture was loaded to a normal phase silica column, the required products remained adsorbed to the column even though a high concentration of methanol was used in the eluent. For the purification process to work a maximum of 20 % of methanol in chloroform was found to be practical. Percentages of methanol can dissolve the silica present in the column. Theoretically, organic solvent mixtures of acetonitrile, ethyl acetate, or even DMF could be used with a higher methanol concentration. I also used preparative and reverse phase HPLC columns to purify very polar compounds. In this study, I obtained the pure compounds **37**, **38** and **41** through this method. Preparative reverse phase HPLC columns suffer from the drawback that the loading is very low. For the mild polar compounds, such as compounds **29** and **30**, I used instead an amine-functionalized silica column (NH-KP, Biotage) to successfully purify the final compounds.

5.5.2.2 Biological activity

Within the squaryldiamide series, three compounds were active: the benzyl-substituted compound **26** as well as the phenyl-substituted ones **27** and **28**. The two phenyl-substituted compounds were found to be the best stabilizers (Table 8). The compounds with a 3-nitro (compound **27**) and a 4-methoxy (compound **28**) substituent have a ΔT_m of 0.78 °C and 0.77 °C, respectively. This is comparable to the T_m -shift observed for ADP (0.69 °C). Among the benzylic compounds, the 4-fluorobenzyl derivative (compound **26**)

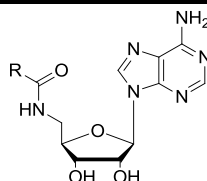
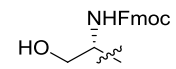
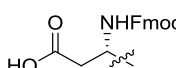
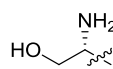
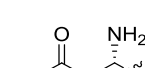
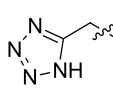
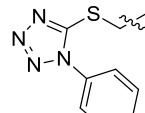
showed a comparable T_m -shift value to AMP (0.47 °C and 0.41 °C, respectively). This indicates that a shorter distance between the adenosine and the aromatic ring is favorable for binding, while additional nitro and methoxy substituents seem to result in equipotent binding. In addition, comparing the structures of the fluoro-substituted benzyl derivative (compound **26**) with the unsubstituted benzyl derivative (compound **23**), compound **23** was not able to stabilize human MDO1. This indicates that activity is gained by the addition of a fluorine atom to the benzylic substituent. The remaining compounds of this series (**24**, **25**, **29** and **30**) did not show stabilization and thus binding. Activity of compound **31** could not be determined due to unspecific high fluorescence in the assay.

Table 8. Docking scores and T_m -shift assay results of human MDO1 for the compounds of the squaryldiamide series.^a

				
Compound	R	Docking score	$T_m/^\circ\text{C}$	$\Delta T_m/^\circ\text{C}$
-	-		47.80	0
AMP	-	-9.16	48.21	0.41
ADP	-	-10.54	48.49	0.69
ADP-ribose	-	-14.80	53.54	5.54
23		-10.53	47.55	-0.25
24		-10.75	47.87	0.07
25		-10.78	47.80	0
26		-12.08	48.27	0.47
27		-9.45	48.58	0.78
28		-9.86	48.57	0.77
29		-10.33	47.89	0.09
30		-8.71	47.85	0.05
31		-12.55	N.D.	N.D.

^a The T_m-shift assay was performed as described in the experimental section and the associated results are expressed as mean values. Compounds were tested at 1 mM concentration. Docking scores were obtained by the XP scoring function (Schrödinger Ltd) as described in the experimental section. The lower the value, the more favorably the ligands bind. N.D. = not determined (unspecific interaction resulting in high fluorescence).

Table 9. Docking scores and T_m-shift assay results of human MDO1 for the compounds of the amide and tetrazole series.^a

				
Compound	R	Docking score	T _m / °C	ΔT _m / °C
34		-9.26	48.38	0.58
36		-11.45	48.55	0.75
37		-8.91	47.66	-0.14
38		-10.47	47.91	0.11
41		-9.88	47.67	-0.13
42		-9.94	47.59	-0.21

^a The T_m-shift assay was performed as described in the experimental section and the associated results are expressed as mean values. Compounds were tested at 1 mM concentration. Docking scores were obtained by the XP scoring function (Schrödinger Ltd) as described in the experimental section. The lower the value, the more favorably the ligands bind.

In the amide series, the Fmoc-protected amino acid derivatives of L-serine (compound **34**) and L-aspartic acid (compound **36**) stabilized the protein, with T_m-shift values of 0.58 °C and 0.75 °C, respectively (Table 2). Only a negligible stabilizing ability of 0.11 °C was observed when the Fmoc group was removed (compound **38**). This means that the large, aromatic and flat tricyclic Fmoc group contributes to binding. Compounds with carboxyl groups (**36** and **38**) clearly retained a better activity than their hydroxy counterparts (**34** and **37**). This indicates that the carboxyl group is a better substituent than hydroxy, likely due to its negative charge.

We could not successfully estimate the K_D for the compounds studied here using isothermal titration calorimetry (ITC, data not shown), perhaps because their activity is weak, in the μM range. The ΔT_m values of the benchmark ligands at 1 mM concentration for human MDO1 are the following: AMP, 0.41 °C; ADP, 0.69 °C; ADPr, 5.54 °C. The ADPr K_D of 0.9 μM has been measured using ITC,(Neuvonen & Ahola, 2009) and ADPr causes the largest T_m -shift in the DSF assay. Similar ΔT_m of 5.98 °C (at 2 mM) and K_D of 5 μM have been reported for ADPr in the viral chikungunya macrodomain protein.(Malet et al., 2009) (The chikungunya macrodomain protein would bind ADP better than human MDO1, with ΔT_m of 4.88 °C for ADP and 0.34 °C for AMP at 2 mM).

5.5.2.3 Docking analysis

We next focused the analysis on the active compounds **26-28**, **34** and **36**, and we hypothesized that these compounds interact with human MDO1 in the same conformational state as ADPr. A preliminary study showed that the structure of MDO1, in the *apo* form, is not suitable for accommodating ADPr, since especially the side chain conformations of Phe306 and Val183 prevent docking. Therefore, a homology model of the human MDO1 protein based on Af1521 was used instead. The binding site in the human MDO1 model could favorably form suitable interactions with ADPr. The binding mode of ADPr extracted from the crystal structure of Af1521 can be successfully reproduced in the structural model of human MDO1 using docking with GlideXP (Figure 38D). The set of compounds was then docked to the model using Glide and the XP scoring function (Tables 8 and 9). ADPr has the best score of all, while compounds **26** and **36** that are good stabilizers have the second and third best scores, respectively. Otherwise there seems to be little correlation between the docking scores and the measured thermal shifts, but docking scores are notoriously poor indicators of binding affinities.

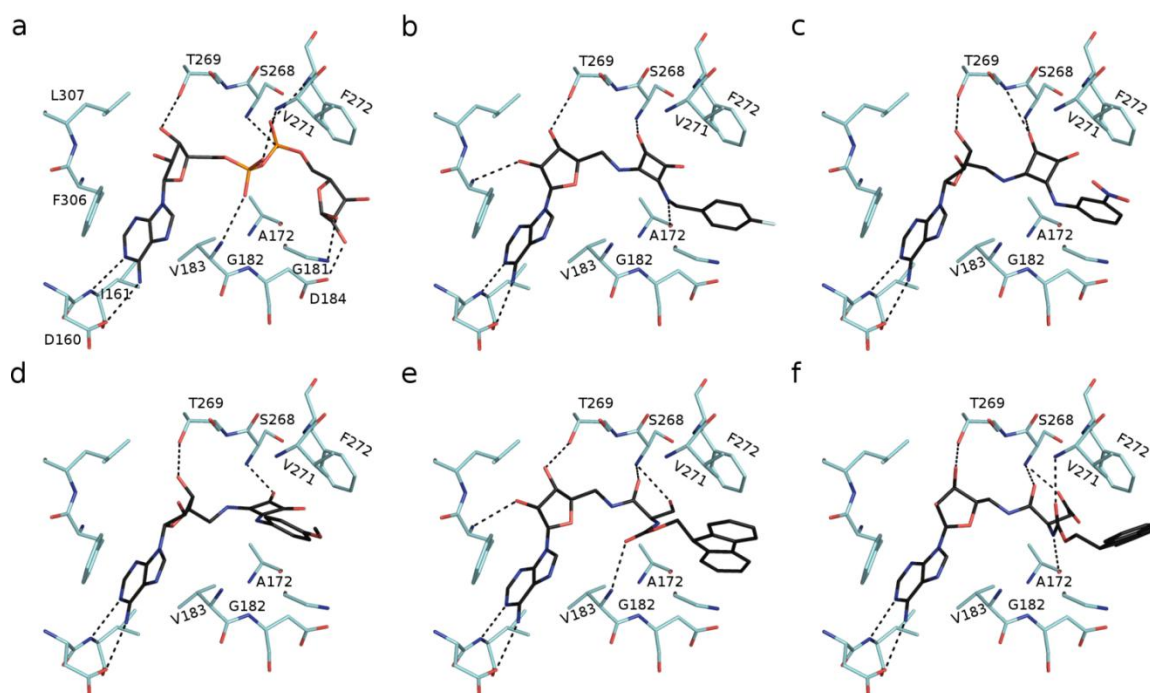


Figure 39. Predicted binding modes of human MDO1 in complex with (a) ADPr, (b) **26**, (c) **27**, (d) **28**, (e) **34** and (f) **36**. Carbon, oxygen, nitrogen and fluorine atoms are shown in black, red, blue, and light blue, respectively. The white cartoon represents the protein, and the carbon atoms of binding site residues are showed in cyan. Putative hydrogen bonds are depicted as dashed lines in green.

Next we will focus on the analysis of the most active compounds **26-28**, **34**, and **36**, considering only the best-ranked pose. For all these compounds, the adenine ring and the ribose occupy the same position as the position predicted for ADPr (Figure 39). The purine ring is sandwiched between the planar phenyl group of Phe224 (π - π interaction) and the aliphatic side-chain of Val183. The NH₂ group of adenine (N6) is donating a hydrogen bond to Asp160. In addition, the N1 nitrogen from the Watson-Crick edge of adenine accepts a hydrogen bond from the main-chain nitrogen of Ile161. The proximal ribose is surrounded by Pro265, Asp310, Cys266 and Thr269. The hydrogen of the ribose O3' hydroxy group donates a hydrogen bond to the side-chain of Thr269.

We next examined which groups are docked in place of the proximal phosphate of ADPr (Figure 39). In MDO1, the negative charge of the proximal phosphate is neutralized by the positively charged macrodipole at the N-terminus of helix 8 (Figure 38b). This macrodipole is generated by the sum of the dipolar contributions from the helical backbone. In addition, on the top of this helix, the main chain NH of Val183 donates a

hydrogen bond to an oxygen atom of the proximal phosphate. Similarly to the phosphate groups of ADPr, the carbonyl of the carbamate group in compound **34** forms a hydrogen bond with the main chain of Val183, while the NH of the carbamate group of compound **36** is in a suitable position to accept a hydrogen bond from the main chain carbonyl of Ala172 (Figures 39 e and f). For compounds 26-28, the two amide NH groups of the squaryldiamide moiety docked approximately at the position of the proximal phosphate of ADPr (Figure 39 b-d). The hydrogen bonding interactions seen in ADPr are however not recreated: in compound **26**, the distal amide NH of the squaryldiamide forms instead a hydrogen bond with the main chain carbonyl of Ala172, whereas the amide hydrogen bond donors do not seem satisfied in compound **27** and **28**. Water molecules do play important role in molecular recognition and may bridge these hydrogen-bonding groups, although they are not included in the docking simulation. On the other hand, it is possible that a next generation of compounds could be designed that contain a hydrogen-bond acceptor pointing towards the main chain NH groups of Val183 and Gly182.

In ADPr, the distal phosphate is hydrogen-bonded to the main chain NH of Ser268, Gly270 and Val271 (Figure 39a, Gly270 is not shown). In compounds **26-28**, one of the squaryldiamide oxygen atoms is located approximately in the same position as one of the oxygen atoms of the distal phosphate, able to hydrogen bond with Ser268 and Gly270, but not Val271. In compounds **34** and **36**, both the carbonyl of the amide linker and the oxygen of the hydroxy and carboxyl groups are well placed to mimic the interactions of the distal phosphate, making hydrogen bonds with the main chain NH of Ser268. Our compounds thus seem to make suitable hydrogen bonds at position equivalent to the distal phosphate.

The third part of ADPr that is mimicked is the distal ribose. In ADPr, the distal ribose is surrounded by Asn174, Asp184, Gly181 and Phe272. The O2' and O3' of ribose form hydrogen bonds to the main chain NH of Gly181 as well as to the side chain of Asp184. In addition, the side chain phenyl of Phe272 is close to the ribose, making hydrophobic interactions with the non-polar edge of the ribose. For compounds **26** and **27**, the nitro- and fluoro-substituted phenyl rings are similarly located; these aryl substituents are thus well placed to make T-shaped stacking with Phe272 as well as edge-to-edge interactions with Asp184. It is possible that the interaction of **28** with Phe272 is less optimal. For compounds **34** and **36** that contain the tricyclic ring of Fmoc, one ring is coming “on top”

of the docked aryl of Phe272 as well as the distal ribose, with the tricyclic ring stacked between Gly181, Val271 and Phe272. Based on these docking results, it seems that the distal ribose can be replaced by aryl moieties and that there is enough space to accommodate a tricyclic ring in the binding pocket of the human MDO1.

5.6 Virtual screening approach

This last section is divided into three parts. (1) Selection of potential compounds to be tested using “functional analogy” (i.e. ADP binding) and PCA for selection. Compounds were purchased from Tocris Bioscience (Bristol, United Kingdom). These are later referred to as TOCRIS compounds (13 molecules). (2) Selection of potential compounds to be tested using structure-based virtual screening. Compounds were purchased from the Chembridge Corporation (San Diego, United States of American). These are later referred to as Chembridge compounds (30 molecules). (3) Experimental testing: three experiments were used: competition assays and ITC and DSC calorimetry.

- **Selection of compounds based on functional similarity**

We hypothesized that compounds binding to an ADP site in a different protein may show cross-binding to MDO1. The hypothesis is that ligands maybe also fit in a macro domain if they fit in ADP site of another protein.

Under this hypothesis, we selected a small dataset of 43 ligands directly from the TOCRIS catalog:

- adenosine receptors ligands (9 agonists and 14 antagonists)
- P2Y receptors (3 agonists and 10 antagonists, 1 unknown)
- 6 methyltransferase inhibitors.

All ligands were put into force field to perform energy-minimization. Ligands with a phosphate and a carboxyl group were deprotonated manually. The structures of the ligands were prepared in Sybylx12.

Principal component analysis (conducted with the software Volsurf+ based on Volsurf+ 122 descriptors) was used to map the chemical space of these compounds. The result is shown in Figure 40. ADPr is shown in green and in the PCA it comes near other adenosine analogs.

Based on PCA analysis of the space, 13 compounds from the Tocris catalog were suggested to be tested for biological activity. The 13 compounds represented by a red dot were selected for experimental testing.

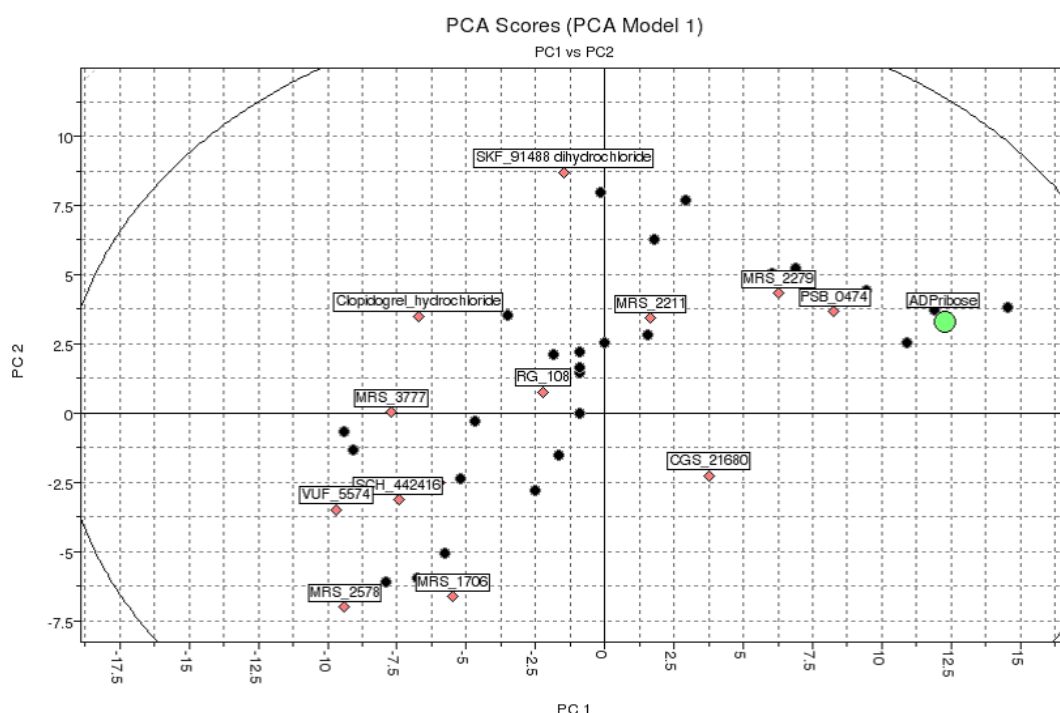


Figure 40. Distribution of the 13 selected ligands among the 43 selected from the TOCRIS catalog (principal component analysis, PC1 and PC2), the selected compounds are marked with a red dot.

- **Selection of compounds based on virtual screening**

The FIMM library, owned by the Institute for Molecular Medicine Finland (FIMM), approximately contains 20,000 lead like compounds. All compounds are available in stock in different amounts. It is our primarily chosen library because it is very convenient to purchase and test compounds from this library. The ligand structures were prepared from 2D .sdf-files to 3D using Balloon (Vainio & Johnson, 2007).

The FIMM library was docked to the *holo* homology structure model of MDO1 (the same model as in Figure 37A) with GOLD 5v2. By default, ten results are available and the

solution with best (highest) score was selected as the most favorable binding mode of ligands. The α -carbon of Val183 was defined as docking center with 20×20×20 grid box. The top 200 of the Gold scores were selected for further analysis. Figure 32 summarizes the docking result. ADPr was used as a control and it obtained the best score; we are very limited because we this is the only control available.

Sixty compounds were chosen manually with visual inspection and in collaboration by a medicinal chemist and based on criteria for example: 1) up to two hydrogen bonds in the adenine binding subpocket; 2) a “bend” in the molecule; 3) the proximal ribose in polar surrounding; 4) hydrogen bond acceptors/donors at the phosphate subpockets. From this set of sixty compounds, we purchased thirty compounds from the Chembridge Corporation taking into account a maximum molecular diversity and the lowest possible price.

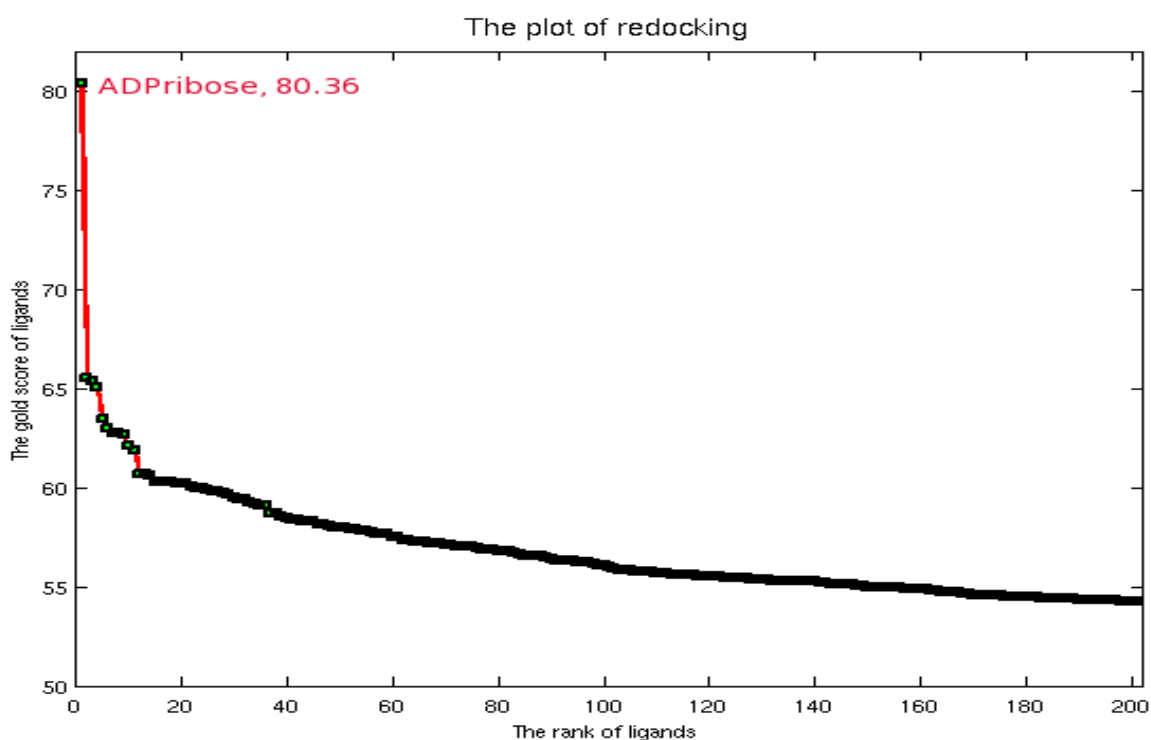


Figure 41. *The Gold docking scores of top 200 FIMM compounds.*

• Experimental testing

Based on the ADPr and PolyADPribose competition assay (Figure 42, Table 10), five compounds from the Tocris set showed inhibition from 41 to 67%. Six compounds from the Chembridge set showed inhibition from 28 to 67% (see Table 10). Poly-ADPr

competition assay was also done for all chosen compounds (see also Table 10). Some of the results do not consist with the results from the ADPr testing. For example, the Tocris candidate MRS 2578 replaced 67% of ADPr, but only 8% of Poly-ADPr. The structures of all active compounds are showed in Figure 43.

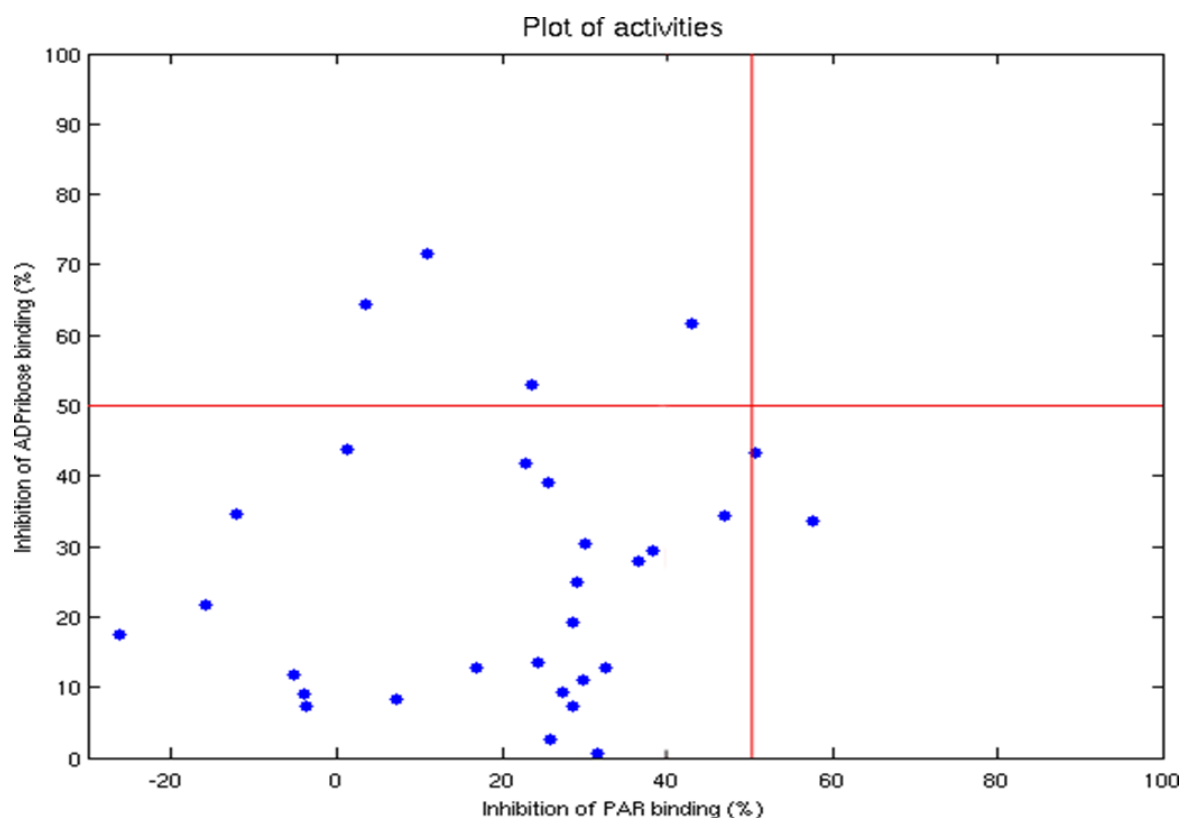


Figure 42. Activities of the 43 compounds tested using ADPr and PAR competition binding assays.

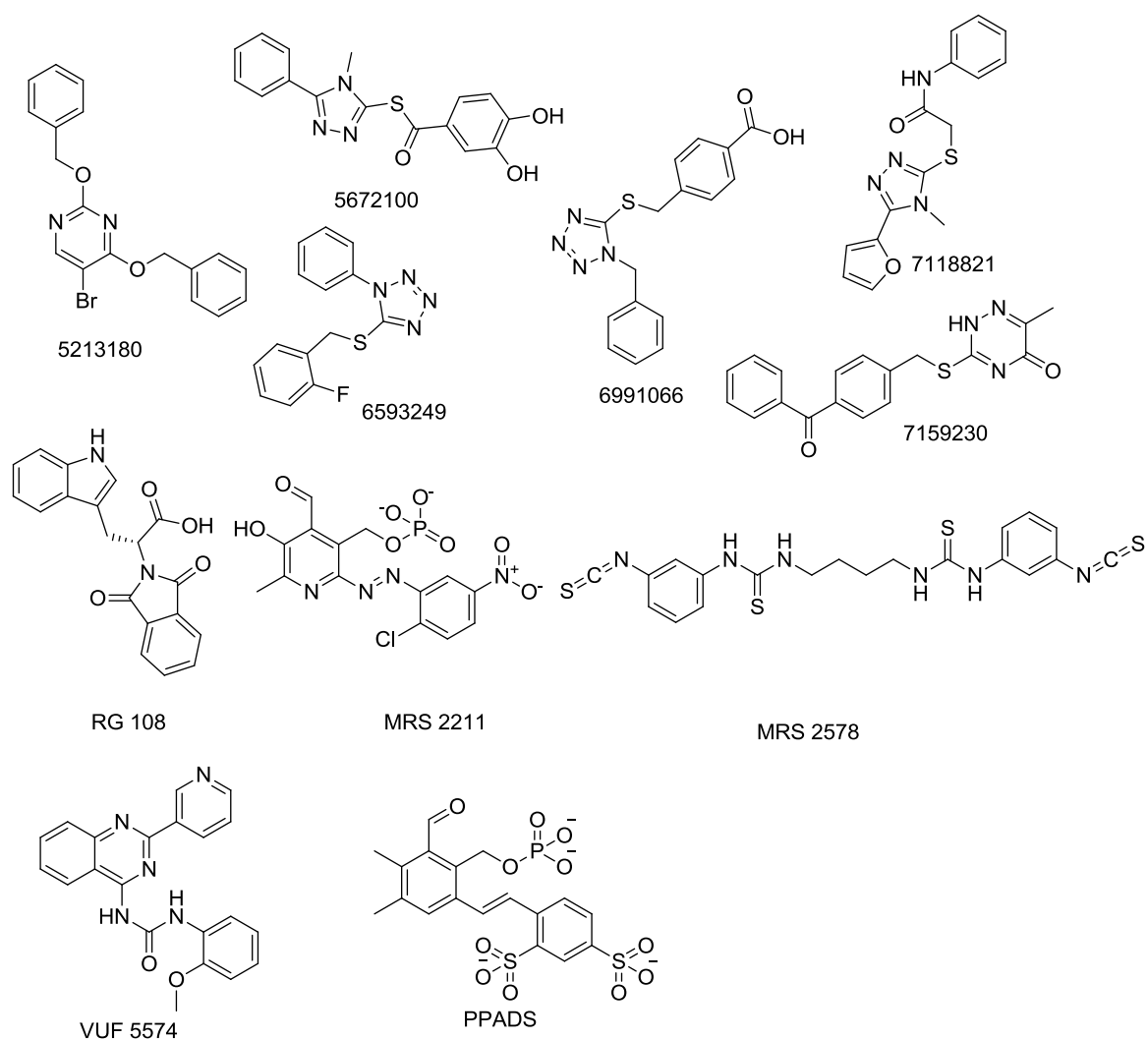


Figure 43. The structures of compounds, which showed activity in the ADPr and poly-ADPr competition assays.

Table 10. *Biological activities of the most active compounds in either ADPr or PAR competition assays.*

Compounds ID	vendor	Percent inhibition (ADP-ribose)(200 nM)	Percent inhibition (poly-ADP-ribose)(20 μ M)	Tested with DSC at 10 μ M	Tested with DSC at 100 μ M
RG 108	Tocris	43	2	Inactive	Inactive
MRS 2211	Tocris	48	96	Inactive	Inactive
MRS 2578	Tocris	67	8	Inactive	Inactive
VUF 5574	Tocris	51	10	Inactive	Inactive
PPADS	Tocris	41	94	-	-
5213180	Chembridge	62	43	Inactive	-
6593249	Chembridge	34	47	Inactive	-
6991066	Chembridge	29	38	-	-
5672011	Chembridge	28	36	-	-
7118821	Chembridge	34	57	Inactive	-
7159230	Chembridge	43	51	Inactive	-

The main conclusion of the experimental testing was that some of the compounds were found active in the ADPr competition assay. However, the assay was not optimized and there were doubts whether the activities were a result of artefacts (ligands inducing the protein to leave its support), leading to false positive results. Furthermore, comparing the results for the ADPr and PAR assays looked inconsistent. Consequently, the compounds were retested using DSC and ITC and at that time, no activity was found at the compound concentration (10-100 μ M) used. During the writing of this thesis, it came out that the compound concentrations used in the assays were too small for detection – at 100 μ M we cannot either detect the ligands synthesized. This led to the decision that the compounds should be retested using a higher concentration, and therefore, we cannot conclusively make a statement about their activities.

6 Perspectives

In this study, the motivation to exemplify the structural replacement of phosphate is as following: 1) phosphate is critical in biological systems; 2) to make our result statistical meaningful, we need enough of data points. In that sense, phosphate is a good candidate because it is widely observed in endogenous ligands. To the best of our knowledge, there are no published studies dealing with the replacement of ribose, while total synthesis of pentose ring is not that easy.

In order to reach medicinal chemists, we have designed a searchable web server. Work is under progress and the website may be functional at the time of defending the thesis. The data generated has been placed on a cloud-based website architecture linked to a Jmol plugin for consultation (written in php Java script an XML), and can also be downloaded from there in part on its in entirety http://86.50.168.121/phosphates_LSR.php (accessible December 4th, 2014).

- **Practical use of the structural replacement study**

Phosphate is a common endogenous substrate in biological systems and a lot of proteins have natural affinity to it. A typical example is pyrophosphatase, which couples pyrophosphate hydrolysis or synthesis to Na⁺ or H⁺ pumping. This protein was the first membrane protein structure ever solved and published in Finland (PDB codes: 4AV3 and 4AV6) (Kellosalo, et al., 2012). It is one main application of my study on structural replacements, and very interesting preliminary results have been obtained. The issue I very soon faced in that project is that chemical libraries useable for virtual screening seldom contain compounds that are focused for a phosphate site, i.e. contain polar and potentially charged groups. Molecules obtained directly from virtual screening did not look promising. I therefore used two strategies for compound selection: the first one was to develop filters that were used to focus the library to fragments identified as isosteres of phosphates, for example the 1,2,5-thiazolidin-3-one-1,1-dioxide. The second was to directly select compounds from vendor catalogs. This second approach has actually revealed some hit compounds with validated binding affinity.

Generalizing and improving the automated workflow

The workflow presented in the thesis could be generalized to exploit structural replacements of other chemical fragments by giving different substructure queries. However, this methodology has limitations; 1) the need of many ligands in the PDB (otherwise we cannot confidently state that one group can replace a target fragment); 2) both the reference and the superimposed protein structure should be in the PDB. For example, tetrazole is a well-known isostere of the carboxyl group while carboxyl groups and phosphates are interchangeable. Thus it can be speculated that tetrazoles can replace phosphates as well, but we did not find such examples simply because there is not enough data points in the PDB.

Another perspective is to develop and generalize the automated workflow to other moieties than phosphate and ribose, for example to peptide/ amide bonds. There are biases in the dataset which have to be taken into consideration, for example we did not consider the individual structural replacement preference against protein family. In our study, it is no surprise that kinases and heat shock proteins dominate the population of our dataset because phosphate is a very common building block of their endogenous ligands. The solution is to cluster proteins into family classes given the protein Uniprot identifier according to Structural Classification of Proteins algorithm. (Andreeva et al., 2008)

Designing better MDO1 ligands

Based on literature as well as our isosteric replacement study of both phosphate and ribose, we discovered many chemical moieties that are able to replace phosphate or ribose in specific circumstances, such as amide and heterocyclic groups. In this thesis work, I have validated the amide and squaryldiamide groups as isosteric replacements of phosphate. Among 15 synthesized adenosine analogs, the binding of four were comparable to ADP and one to AMP.

We explored the substitution of the phosphate-distal ribose part of ADPr. The next step would be to substitute the adenine-proximal ribose part of ADPr, for example using potential ribose replacements found in this thesis. We have not tested our findings of

ribose isosteric replacement. As a continuation of this project, it would be worth combining both the phosphate and ribose isosteric replacements we have found. Subsequently, novel compounds differing from ADP could be developed. Even if it has not been tested, our ligands could be speculated to bind to other proteins. Changing the adenine part should also improve the specificity of our ligands towards MDO1.

Another area is to improve the potency of our ligands. Based on docking, new derivatives can be synthesized to further optimize the protein-ligand interactions. For example, an interaction with helix 8 can be an option since the α -helical dipole putatively exists (Sengupta et al., 2005) and it is critical to phosphate binding, but there is no hydrogen-bond acceptor in my compounds. We could for example develop a compound with a cyclic moiety, e.g. a cyclopentane-1,3-dione, with two branched carbonyl that may form a backbone hydrogen bond of both helix 8 and turn 7. In addition, more optimization can be performed on the most distal part of the molecule. Here we could take advantage of the flexibility of the turn 8 and to prepare compounds with substituents of different sizes. This would allow us to develop compounds that do not occupy the adenine-proximal ribose binding site, since we do not want to make very large molecules in the end.

What's more, we have discovered a potential binding location that can favorably accept a tricyclic rings. We could now try to place for example fluorescein to develop a fluorescent probe useful for screening. According to our collaborator, even a low-potency probe would be useful. It was surprising and counter-intuitive that Fmoc was accommodated by the binding site since it is a large substituent.

Retesting the virtual screening hits with thermofluor

The results of the virtual screening study I performed with the purpose of finding new ligands for human MDO1 (see chapter V.6.) are not conclusive so far; activity of the compounds was found by the competition assay but no activity by the preliminary calorimetric study. However, the concentration of the compounds used for testing in the calorimetric method was outside of the detection window, i.e. they should have been in the mM range instead of in the μ M range. Thus, these molecules should be retested at mM concentrations.

Concluding remarks

As a general perspective, my study not only validated the variety of local structural replacements, but also suggested additional possibilities. In addition, the conditionality of these surrogates is emphasized, since replacements are circumstance (binding-site) dependent. One isosteric replacement validated in one binding site cannot directly be translated into another. In terms of the synthesized compounds, the ones containing a carboxyl group were surprisingly found not to increase the binding to MDO1. The squaryldiamide and amide groups were both useful isosteres of phosphate. The most surprising is that the major determinant of potency was found in the presence of hydrophobic groups, i.e. the compounds containing a phenyl or benzyl group and the presence of the Fmoc group. Our study thus emphasized the versatility of mechanisms that accompanies structural replacements; they all blur the definition of “structural similarity” presented in the Review of Literature and as the “Structure-Property-Principle”, since these two concepts are based in the *conservation* of molecular interactions.

References

- Abel, R., Young, T., Farid, R., Berne, B. J., & Friesner, R. A. (2008). Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *Journal of the American Chemical Society*, 130(9), 2817–31.
- Allen, F. H., Groom, C. R., Liebeschuetz, J. W., Bardwell, D. a, Olsson, T. S. G., & Wood, P. a. (2012). The hydrogen bond environments of 1H-tetrazole and tetrazolate rings: the structural basis for tetrazole-carboxylic acid bioisosterism. *Journal of Chemical Information and Modeling*, 52(3), 857–66.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410.
- Amadasi, A., Surface, J. A., Spyraakis, F., Cozzini, P., Mozzarelli, A., & Kellogg, G. E. (2008). Robust classification of “relevant” water molecules in putative protein binding sites. *Journal of Medicinal Chemistry*, 51, 1063–1067.
- Andersson, H., Demaegdt, H., Johnsson, A., Vauquelin, G., Lindeberg, G., Hallberg, M., ... Hallberg, A. (2011). Potent macrocyclic inhibitors of insulin-regulated aminopeptidase (IRAP) by olefin ring-closing metathesis. *Journal of Medicinal Chemistry*, 54(11), 3779–92.
- Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*, 36(Database issue), D419–25.
- Ayesa, S., Samuelsson, B., & Classon, B. (2008). A One-Pot, Solid-Phase Synthesis of Secondary Amines from Reactive Alkyl Halides and an Alkyl Azide. *Synlett*, 2008(1), 97–99. 7
- Azam, M., Seeliger, M. A., Gray, N. S., Kuriyan, J., & Daley, G. Q. (2008). Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nature Structural & Molecular Biology*, 15(10), 1109–18.
- Bagriantsev, S. N., Ang, K. H., Gallardo-Godoy, A., Clark, K. A., Arkin, M. R., Renslo, A. R., & Minor, D. L. (2013). A high-throughput functional screen identifies small molecule regulators of temperature- and mechano-sensitive K2P channels. *ACS Chemical Biology*, 8, 1841–1851.
- Ballatore, C., Huryn, D. M., & Smith, A. B. (2013). Carboxylic acid (bio)isosteres in drug design. *ChemMedChem*, 8(3), 385–95.
- Bao, Z. Q., Jacobsen, D. M., & Young, M. A. (2011). Briefly bound to activate: transient binding of a second catalytic magnesium activates the structure and dynamics of CDK2 kinase for catalysis. *Structure (London, England : 1993)*, 19(5), 675–90.
- Berman, H. M. (2000a). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242.
- Berman, H. M. (2000b). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242.
- Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., ... Schwede, T. (2014a). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(Web Server issue),
- Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., ... Schwede, T. (2014b). SWISS-MODEL: modelling protein tertiary and quaternary structure

- using evolutionary information. *Nucleic Acids Research*, 42(Web Server issue), W252–8.
- Biela, A., Nasief, N. N., Betz, M., Heine, A., Hangauer, D., & Klebe, G. (2013). Dissecting the hydrophobic effect on the molecular level: the role of water, enthalpy, and entropy in ligand binding to thermolysin. *Angewandte Chemie (International Ed. in English)*, 52(6), 1822–8.
- Biot, C., & Bauer, H. (2004). 5-substituted tetrazoles as bioisosteres of carboxylic acids. Bioisosterism and mechanistic studies on glutathione reductase inhibitors as antimalarials. *Journal of Medicinal Chemistry*, 47(24), 5972–83.
- Biot, C., Dessolin, J., Grellier, P., & Davioud-Charvet, E. (2003). Double-drug development against antioxidant enzymes from *Plasmodium falciparum*. *Redox Report : Communications in Free Radical Research*, 8(5), 280–3.
- Bissantz, C., Kuhn, B., & Stahl, M. (2010). A medicinal chemist's guide to molecular interactions. *Journal of Medicinal Chemistry*, 53(14), 5061–84.
- Boström, J., Hogner, A., & Schmitt, S. (2006). Do structurally similar ligands bind in a similar fashion? *Journal of Medicinal Chemistry*, 49, 6716–6725.
- Bross, P. F., Kane, R., Farrell, A. T., Abraham, S., Benson, K., Brower, M. E., ... Pazdur, R. (2004). Approval summary for bortezomib for injection in the treatment of multiple myeloma. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, 10(12 Pt 1), 3954–64.
- Bruno, I. J., Cole, J. C., Lommerse, J. P., Rowland, R. S., Taylor, R., & Verdonk, M. L. (1997). IsoStar: a library of information about nonbonded interactions. *Journal of Computer-Aided Molecular Design*, 11(1), 525–537.
- Burger, A. (1991). Isosterism and bioisosterism in drug design. *Progress in Drug Research. Fortschritte Der Arzneimittelforschung. Progrès Des Recherches Pharmaceutiques*, 37, 287–371.
- Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/1763185>
- Böhm, H. J., Flohr, A., & Stahl, M. (2004). Scaffold hopping. *Drug Discovery Today: Technologies*, 1, 217–224.
- Carini, D. J., Duncia, J. V., Aldrich, P. E., Chiu, A. T., Johnson, A. L., Pierce, M. E., ... Wells, G. J. (1991). Nonpeptide angiotensin II receptor antagonists: the discovery of a series of N-(biphenylmethyl)imidazoles as potent, orally active antihypertensives. *Journal of Medicinal Chemistry*, 34(8), 2525–2547.
- Chen, D., Vollmar, M., Rossi, M. N., Phillips, C., Kraehenbuehl, R., Slade, D., ... Ahel, I. (2011). Identification of macrodomain proteins as novel O-acetyl-ADP-ribose deacetylases. *The Journal of Biological Chemistry*, 286(15), 13261–13271.
- Chikhi, A., & Bensegueni, A. (2008). Docking Efficiency Comparison of Surflex, a Commercial Package and Arguslab, a Licensable Freeware. *Journal of Computer Science & Systems Biology*, 01(01), 081–086.
- Chodera, J. D., & Mobley, D. L. (2013). Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annual Review of Biophysics*, 42, 121–42.
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11), 1422–3.

- Cruciani, G., Crivori, P., Carrupt, P., & Testa, B. (2000). Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *Journal of Molecular Structure*, 503(1-2), 17–30.
- Cumming, J. G., Davis, A. M., Muresan, S., Haeblerlein, M., & Chen, H. (2013). Chemical predictive modelling to improve compound quality. *Nature Reviews. Drug Discovery*, 12(12), 948–62.
- Dalgarno, D., Stehle, T., Narula, S., Schelling, P., van Schravendijk, M. R., Adams, S., ... Sawyer, T. (2006). Structural basis of Src tyrosine kinase inhibition with a new class of potent and selective trisubstituted purine-based compounds. *Chemical Biology & Drug Design*, 67(1), 46–57.
- Davioud-Charvet, E., Delarue, S., Biot, C., Schwöbel, B., Boehme, C. C., Müssigbrodt, A., ... Becker, K. (2001). A Prodrug Form of a Plasmodium falciparum Glutathione Reductase Inhibitor Conjugated with a 4-Anilinoquinoline. *Journal of Medicinal Chemistry*, 44(24), 4268–4276.
- Denesyuk, A., & Denessiouk, K. (2003). Phosphate group binding “cup” of PLP-dependent and non-PLP-dependent enzymes: leitmotif and variations. ... *et Biophysica Acta (BBA ...)*, 1647(1-2), 234–238.
- Desaphy, J., & Rognan, D. (2014). sc-PDB-Frag: a database of protein-ligand interaction patterns for Bioisosteric replacements. *Journal of Chemical Information and Modeling*, 54(7), 1908–18.
- DeWitt, D. L. (1999). Cox-2-selective inhibitors: the new super aspirins. *Molecular Pharmacology*, 55(4), 625–631.
- Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The protein folding problem. *Annual Review of Biophysics*, 37, 289–316.
- Dimova, D., Heikamp, K., Stumpfe, D., & Bajorath, J. (2013). Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets. *Journal of Medicinal Chemistry*, 56(8), 3339–45.
- Dollins, D. E., Warren, J. J., Immormino, R. M., & Gewirth, D. T. (2007). Structures of GRP94-nucleotide complexes reveal mechanistic differences between the hsp90 chaperones. *Molecular Cell*, 28(1), 41–56.
- Dong, L., Marakovits, J., Hou, X., Guo, C., Greasley, S., Dagostino, E., ... Murray, B. W. (2010). Structure-based design of novel human Pin1 inhibitors (II). *Bioorganic & Medicinal Chemistry Letters*, 20(7), 2210–4.
- Dougherty, D. A. (2007). Cation- π Interactions Involving Aromatic Amino Acids. *The Journal of Nutrition*, 137(6), 1504S–1508S.
- Dougherty, D. A. (2013). The cation- π interaction. *Accounts of Chemical Research*, 46, 885–893.
- Doweyko, A. M. (2008). QSAR: dead or alive? *Journal of Computer-Aided Molecular Design*, 22(2), 81–9.
- Downey, a. M., & Cairo, C. W. (2014). Synthesis of α -brominated phosphonates and their application as phosphate bioisosteres. *Med. Chem. Commun.*, 00, 1–15.
- Du, Q.-S., Wang, Q.-Y., Du, L.-Q., Chen, D., & Huang, R.-B. (2013). Theoretical study on the polar hydrogen- π (Hp- π) interactions between protein side chains. *Chemistry Central Journal*, 7(1), 92–99.

- DuBois, R. M., Slavish, P. J., Baughman, B. M., Yun, M.-K., Bao, J., Webby, R. J., ... White, S. W. (2012). Structural and biochemical basis for development of influenza virus inhibitors targeting the PA endonuclease. *PLoS Pathogens*, 8(8), e1002830.
- Elliott, T. S., Slowey, A., Ye, Y., & Conway, S. J. (2012). The use of phosphate bioisosteres in medicinal chemistry and chemical biology. *MedChemComm*, 3(7), 735.
- Ellis, G. A., Palte, M. J., & Raines, R. T. (2012). Boronate-mediated biologic delivery. *Journal of the American Chemical Society*, 134, 3631–3634.
- Ertl, P. (2012). Database of bioactive ring systems with calculated properties and its use in bioisosteric design and scaffold hopping. *Bioorganic & Medicinal Chemistry*, 20(18), 5436–42.
- Eugene Kellogg, G., & Abraham, D. J. (2000). Hydrophobicity: is LogPo/w more than the sum of its parts? *European Journal of Medicinal Chemistry*, 35(7-8), 651–661.
- Fanelli, R., Schembri, L., Piarulli, U., Pinoli, M., Rasini, E., Paolillo, M., ... Marino, F. (2014). Effects of a novel cyclic RGD peptidomimetic on cell proliferation, migration and angiogenic activity in human endothelial cells. *Vascular Cell*, 6, 11–19.
- FitzGerald, G. A., & Patrono, C. (2001). The coxibs, selective inhibitors of cyclooxygenase-2. *The New England Journal of Medicine*, 345(6), 433–42.
- Floris, M., Masciocchi, J., Fanton, M., & Moro, S. (2011). Swimming into peptidomimetic chemical space using pepMMsMIMIC. *Nucleic Acids Research*, 39.
- Floris, M., & Moro, S. (2012). Mimicking peptides??? in silico. *Molecular Informatics*, 31, 12–20.
- Flower, R. J. (2003). The development of COX2 inhibitors. *Nature Reviews. Drug Discovery*, 2(3), 179–91.
- Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., ... Mainz, D. T. (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21), 6177–96.
- Gajewiak, J., & Prestwich, G. D. (2006). Phosphomimetic sulfonamide and sulfonamidoxo analogues of (Lyso)phosphatidic acid. *Tetrahedron Letters*, 47, 7607–7609.
- Gallivan, J. P., & Dougherty, D. A. (2000). A computational study of cation-?? interactions vs salt bridges in aqueous media: Implications for protein engineering. *Journal of the American Chemical Society*, 122, 870–874.
- Ganellin, C. R. (1977). Relative concentrations of zwitterionic and uncharged species in catecholamines and the effect of N-substituents. *Journal of Medicinal Chemistry*, 20, 579–581.
- Gans, K. R., Galbraith, W., Roman, R. J., Haber, S. B., Kerr, J. S., Schmidt, W. K., Ackerman, N. R. (1990). Anti-inflammatory and safety profile of DuP 697, a novel orally effective prostaglandin synthesis inhibitor. *The Journal of Pharmacology and Experimental Therapeutics*, 254(1), 180–187.
- García-Sosa, A. T. (2013). Hydration properties of ligands and drugs in protein binding sites: tightly-bound, bridging water molecules and their effects and consequences on molecular design strategies. *Journal of Chemical Information and Modeling*, 53(6), 1388–405.

- García-Sosa, A. T., Mancera, R. L., & Dean, P. M. (2003). WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *Journal of Molecular Modeling*, 9(3), 172–82.
- George, P., & Nathan, B. (2013). In silico applications of bioisosterism in contemporary medicinal chemistry practice. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(4), 339–354.
- Gilli, P., Pretto, L., Bertolasi, V., & Gilli, G. (2009). Predicting Hydrogen-Bond strengths from Acid-Base molecular properties. the pKa slide rule: Toward the solution of a Long-Lasting problem. *Accounts of Chemical Research*, 42, 33–44.
- Goetz, G. H., Farrell, W., Shalaeva, M., Sciabola, S., Anderson, D., Yan, J., ... Shapiro, M. J. (2014). High throughput method for the indirect detection of intramolecular hydrogen bonding. *Journal of Medicinal Chemistry*, 57, 2920–2929.
- Golbraikh, A., Muratov, E., Fourches, D., & Tropsha, A. (2014). Data set modelability by QSAR. *Journal of Chemical Information and Modeling*, 54, 1–4.
- Golbraikh, A., & Tropsha, A. (2002). Beware of q²! *Journal of Molecular Graphics and Modelling*, 20(4), 269–276.
- Hall, B. Y. H. K. (1957). RATES OF AMIDATION Sterically Hindered Phenolic Buffers . Application to Determination of Rates of Amidation of Ethyl Chloroformate '. *J. Am. Chem. Soc.*, 79(20), 5439–5441.
- Han, W., Li, X., & Fu, X. (2011). The macro domain protein family: structure, functions, and their potential therapeutic implications. *Mutation Research*, 727(3), 86–103.
- Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T. M., Mortenson, P. N., & Murray, C. W. (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry*, 50(4), 726–41.
- Hawkins, P. C. D., Skillman, A. G., & Nicholls, A. (2007). Comparison of shape-matching and docking as virtual screening tools. *Journal of Medicinal Chemistry*, 50, 74–82.
- Hayward, J. (2012). Part Two Data. In *Bioisosteres in Medicinal Chemistry* (pp. 53–74).
- He, J., Gajewiak, J., Scott, J. L., Gong, D., Ali, M., Best, M. D., ... Kutateladze, T. G. (2011). Metabolically stabilized derivatives of phosphatidylinositol 4-phosphate: Synthesis and applications. *Chemistry and Biology*, 18, 1312–1319.
- He, Y., Wilkins, J. P., & Kiessling, L. L. (2006). N-acylsulfonamide linker activation by Pd-catalyzed allylation. *Organic Letters*, 8(12), 2483–5.
- Hess, M., Schulze Elfringhoff, A., & Lehr, M. (2007). 1-(5-Carboxy- and 5-carbamoylindol-1-yl)propan-2-ones as inhibitors of human cytosolic phospholipase A2alpha: bioisosteric replacement of the carboxylic acid and carboxamide moiety. *Bioorganic & Medicinal Chemistry*, 15(8), 2883–91.
- Hikaru Harada. (1971). An Investigation of the Stability Constant of the 2,3-Dihydroxybenzoic Acid Complex with Copper(II). *Bulletin of the Chemical Society of Japan*, 44(12), 3459–3460.
- Hu, B., & Lill, M. a. (2014). WATsite: hydration site prediction program with PyMOL interface. *Journal of Computational Chemistry*, 35(16), 1255–60.

- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9, 99–104.
- Hussain, M., Ahmed, V., Hill, B., Ahmed, Z., & Taylor, S. D. (2008). A re-examination of the difluoromethylenesulfonic acid group as a phosphotyrosine mimic for PTP1B inhibition. *Bioorganic and Medicinal Chemistry*, 16, 6764–6777.
- Högenauer, K., Hinterding, K., & Nussbaumer, P. (2010). S1P receptor mediated activity of FTY720 phosphate mimics. *Bioorganic & Medicinal Chemistry Letters*, 20(5), 1485–7.
- Imagama, S., Abe, A., Suzuki, M., Hayakawa, F., Katsumi, A., Emi, N., ... Naoe, T. (2007). LRP16 is fused to RUNX1 in monocytic leukemia cell line with t(11;21)(q13;q22). *European Journal of Haematology*, 79(1), 25–31.
- Immormino, R. M., Dollins, D. E., Shaffer, P. L., Soldano, K. L., Walker, M. A., & Gewirth, D. T. (2004). Ligand-induced conformational shift in the N-terminal domain of GRP94, an Hsp90 chaperone. *The Journal of Biological Chemistry*, 279(44), 46162–71.
- Immormino, R. M., Metzger, L. E., Reardon, P. N., Dollins, D. E., Blagg, B. S. J., & Gewirth, D. T. (2009). Different poses for ligand and chaperone in inhibitor-bound Hsp90 and GRP94: implications for paralog-specific drug design. *Journal of Molecular Biology*, 388(5), 1033–42.
- Jackson, M. R., Beahm, R., Duvvuru, S., Narasimhan, C., Wu, J., Wang, H. N., ... Howell, E. E. (2007). A preference for edgewise interactions between aromatic rings and carboxylate anions: The biological relevance of anion-quadrupole interactions. *Journal of Physical Chemistry B*, 111, 8242–8249.
- Jankevicius, G., Hassler, M., Golia, B., Rybin, V., Zacharias, M., Timinszky, G., & Ladurner, A. G. (2013). A family of macrodomain proteins reverses cellular mono-ADP-ribosylation. *Nature Structural & Molecular Biology*, 20(4), 508–14.
- Jinhua, Z., Kleinöder, T., & Gasteiger, J. (2006). Prediction of pKa values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. *Journal of Chemical Information and Modeling*, 46, 2256–2266.
- Johnson, M., Zaretskaya, I., Raytselis, Y., Merezuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36(Web Server issue), W5–9.
- Jorgensen, W. L. (1991). Rusting of the lock and key model for protein-ligand binding. *Science (New York, N.Y.)*, 254, 954–955.
- Karras, G. I., Kustatscher, G., Buhecha, H. R., Allen, M. D., Pugieux, C., Sait, F., ... Ladurner, A. G. (2005). The macro domain is an ADP-ribose binding module. *The EMBO Journal*, 24(11), 1911–20.
- Kellogg, G. E., & Chen, D. L. (2004). The importance of being exhaustive. Optimization of bridging structural water molecules and water networks in models of biological systems. *Chemistry & Biodiversity*, 1(1), 98–105.
- Kellosalo, J., Kajander, T., Kogan, K., Pokharel, K., & Goldman, A. (2012). The structure and catalytic cycle of a sodium-pumping pyrophosphatase. *Science (New York, N.Y.)*, 337(6093), 473–6.
- Kramer, C., & Fuchs, J. (2014). Matched Molecular Pair Analysis: Significance and the Impact of Experimental Uncertainty. *Journal of Medicinal ...*, 57(9), 3786–3802.

- Kuhn, B., Mohr, P., & Stahl, M. (2010). Intramolecular hydrogen bonding in medicinal chemistry. *Journal of Medicinal Chemistry*, 53, 2601–2611.
- Kumar, A., Mandiyan, V., Suzuki, Y., Zhang, C., Rice, J., Tsai, J., Bremer, R. (2005). Crystal structures of proto-oncogene kinase Pim1: a target of aberrant somatic hypermutations in diffuse large cell lymphoma. *Journal of Molecular Biology*, 348(1), 183–93.
- Labby, K. J., Xue, F., Kraus, J. M., Ji, H., Mataka, J., Li, H., ... Silverman, R. B. (2012). Intramolecular hydrogen bonding: A potential strategy for more bioavailable inhibitors of neuronal nitric oxide synthase. *Bioorganic and Medicinal Chemistry*, 20, 2435–2443.
- Lajiness, M. S., Maggiora, G. M., & Shanmugasundaram, V. (2004). Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *Journal of Medicinal Chemistry*, 47(20):4891–6
- Lam, P., Jadhav, P., Eyermann, C., Hodge, C., Ru, Y., Bacheler, L., ... Et, A. (1994). Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science*, 263(5145), 380–384.
- Langdon, S. R., Ertl, P., & Brown, N. (2010). Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Molecular Informatics*.
- Langmuir, I. (1919). Isomorphism, isosterism and covalence. *Journal of the American Chemical Society*, 41, 1543–1559.
- Lao, B. B., Grishagin, I., Mesallati, H., Brewer, T. F., Olenyuk, B. Z., & Arora, P. S. (2014). In vivo modulation of hypoxia-inducible signaling by topographical helix mimetics. *Proceedings of the National Academy of Sciences of the United States of America*, 8–11.
- Leach, A., & Jones, H. (2006). Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *Journal of Medicinal Chemistry*, 49(23), 6672–6682.
- Li, Q., Woods, K. W., Thomas, S., Zhu, G.-D., Packard, G., Fisher, J., ... Giranda, V. L. (2006). Synthesis and structure-activity relationship of 3,4'-bispyridinylethylenes: discovery of a potent 3-isoquinolinylnpyridine inhibitor of protein kinase B (PKB/Akt) for the treatment of cancer. *Bioorganic & Medicinal Chemistry Letters*, 16, 2000–2007.
- Li, Z., He, Y., Cao, L., Wong, L., & Li, J. (2012). Conservation of water molecules in protein binding interfaces. *International Journal of Bioinformatics Research and Applications*, 8(3-4), 228–44. 68
- Liedtke, A. J., Crews, B. C., Daniel, C. M., Blobaum, A. L., Kingsley, P. J., Ghebreselasie, K., & Marnett, L. J. (2012). Cyclooxygenase-1-selective inhibitors based on the (E)-2'-des-methyl-sulindac sulfide scaffold. *Journal of Medicinal Chemistry*, 55, 2287–300.
- Liu, D., Bimbo, L. M., Mäkilä E., Villanova, F., Kaasalainen, M., Herranz-Blanco, B., ... Santos, H. A. (2013). Co-delivery of a hydrophobic small molecule and a hydrophilic peptide by porous silicon nanoparticles. *Journal of Controlled Release*, 170, 268–278.
- Lloyd, D. G., García-Sosa, A. T., Alberts, I. L., Todorov, N. P., & Mancera, R. L. (2004). The effect of tightly bound water molecules on the structural interpretation of ligand-

- derived pharmacophore models. *Journal of Computer-Aided Molecular Design*, 18(2), 89–100.
- Locock, K. E. S., Yamamoto, I., Tran, P., Hanrahan, J. R., Chebib, M., Johnston, G. A. R., & Allan, R. D. (2013). γ -Aminobutyric Acid(C) (GABAC) Selective Antagonists Derived from the Bioisosteric Modification of 4-Aminocyclopent-1-enecarboxylic Acid: Amides and Hydroxamates. *Journal of Medicinal Chemistry*, 56(13), 5626–5630.
- Lolli, M. L., Giordano, C., Pickering, D. S., Rolando, B., Hansen, K. B., Foti, A., ... Johansen, T. N. (2010). 4-Hydroxy-1,2,5-oxadiazol-3-yl moiety as bioisoster of the carboxy function. Synthesis, ionization constants, and molecular pharmacological characterization at ionotropic glutamate receptors of compounds related to glutamate and its homologues. *Journal of Medicinal Chemistry*, 53, 4110–4118.
- Lolli, M. L., Hansen, S. L., Rolando, B., Nielsen, B., Wellendorph, P., Madsen, K., ... Johansen, T. N. (2006). Hydroxy-1,2,5-oxadiazolyl moiety as bioisoster of the carboxy function. Synthesis, ionization constants, and pharmacological characterization of γ -aminobutyric acid (GABA) related compounds. *Journal of Medicinal Chemistry*, 49, 4442–4446.
- Lummis, S. C. R., Beene, D. L., Harrison, N. J., Lester, H. A., & Dougherty, D. A. (2005). A cation- π binding interaction with a tyrosine in the binding site of the GABAC receptor. *Chemistry and Biology*, 12, 993–997.
- Madhusudan, Trafny, E. A., Xuong, N. H., Adams, J. A., Ten Eyck, L. F., Taylor, S. S., & Sowadski, J. M. (1994). cAMP-dependent protein kinase: crystallographic insights into substrate recognition and phosphotransfer. *Protein Science : A Publication of the Protein Society*, 3(2), 176–87.
- Maggiora, G. M. (2006). On outliers and activity cliffs--why QSAR often disappoints. *Journal of Chemical Information and Modeling*, 46(4), 1535–1535.
- Maggiora, G., Vogt, M., Stumpfe, D., & Bajorath, J. (2013). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8), 3186–204.
- Malet, H., Coutard, B., Jamal, S., Dutartre, H., Papageorgiou, N., Neuvonen, M., ... Canard, B. (2009). The crystal structures of Chikungunya and Venezuelan equine encephalitis virus nsP3 macro domains define a conserved adenosine binding pocket. *Journal of Virology*, 83(13), 6534–45.
- Malik, M. A., Wani, M. Y., Al-Thabaiti, S. A., & Shiekh, R. A. (2014). Tetrazoles as carboxylic acid isosteres: chemistry and biology. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*, 78(1-4), 15–37.
- Manning, G., Whyte, D. B., Martinez, R., Hunter, T., & Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science (New York, N.Y.)*, 298(5600), 1912–34.
- Martinez, C. R., & Iverson, B. L. (2012). Rethinking the term “ π -stacking.” *Chemical Science*.
- McGaughey, G. B. (1998). π -Stacking Interactions. ALIVE AND WELL IN PROTEINS. *Journal of Biological Chemistry*, 273(25), 15458–15463.
- Meanwell, N. A. (2011). Synopsis of some recent tactical application of bioisosteres in drug design. *Journal of Medicinal Chemistry*, 54(8), 2529–91.

- Menegatti, S., Hussain, M., Naik, A. D., Carbonell, R. G., & Rao, B. M. (2013). mRNA display selection and solid-phase synthesis of Fc-binding cyclic peptide affinity ligands. *Biotechnology and Bioengineering*, 110, 857–870. 0
- Meng, X.-Y., Zhang, H.-X., Mezei, M., & Cui, M. (2011). Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design*, 7(2), 146–157.
- Michel, J., Tirado-Rives, J., & Jorgensen, W. L. (2009). Prediction of the water content in protein binding sites. *The Journal of Physical Chemistry. B*, 113, 13337–13346.
- Moreau, C., Kirchberger, T., Swarbrick, J. M., Bartlett, S. J., Fliegert, R., Yorgan, T., ... Potter, B. V. L. (2013). Structure-activity relationship of adenosine 5'-diphosphoribose at the transient receptor potential melastatin 2 (TRPM2) channel: rational design of antagonists. *Journal of Medicinal Chemistry*, 56(24), 10079–102.
- Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16), 2785–91.
- Neuvonen, M., & Ahola, T. (2009). Differential activities of cellular and viral macro domain proteins in binding of ADP-ribose metabolites. *Journal of Molecular Biology*, 385(1), 212–25.
- Niewiadomski, S., Beebejaun, Z., Denton, H., Smith, T. K., Morris, R. J., & Wagner, G. K. (2010). Rationally designed squaryldiamides - a novel class of sugar-nucleotide mimics? *Organic & Biomolecular Chemistry*, 8(15), 3488–99.
- O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An Open chemical toolbox. *Journal of Cheminformatics*, 3.
- Olsson, T. S. G., Williams, M. a, Pitt, W. R., & Ladbury, J. E. (2008). The thermodynamics of protein-ligand interaction and solvation: insights for ligand design. *Journal of Molecular Biology*, 384(4), 1002–17.
- Ozawa, T., Tsuji, E., Ozawa, M., Handa, C., Mukaiyama, H., Nishimura, T., ... Okazaki, K. (2008). The importance of CH/pi hydrogen bonds in rational drug design: An ab initio fragment molecular orbital study to leukocyte-specific protein tyrosine (LCK) kinase. *Bioorganic & Medicinal Chemistry*, 16(24), 10311–8.
- Pajouhesh, H., & Lenz, G. R. (2005). Medicinal chemical properties of successful central nervous system drugs. *NeuroRx: The Journal of the American Society for Experimental NeuroTherapeutics*, 2(4), 541–53.
- Philip, V., Harris, J., Adams, R., Nguyen, D., Spiers, J., Baudry, J., Hinde, R. J. (2011). A Survey of Aspartate À Phenylalanine and Glutamate À Phenylalanine. *Biochemistry*, 50(14), 2939–2950.
- Politzer, P., Lane, P., Concha, M. C., Ma, Y., & Murray, J. S. (2007). An overview of halogen bonding. *Journal of Molecular Modeling*.
- Politzer, P., Murray, J. S., & Clark, T. (2013). Halogen bonding and other σ -hole interactions: a perspective. *Physical Chemistry Chemical Physics: PCCP*, 15(27), 11178–89.
- Qian, K., Yu, D., Chen, C.-H., Huang, L., Morris-Natschke, S. L., Nitz, T. J., ... Lee, K.-H. (2009). Anti-AIDS agents. 78. Design, synthesis, metabolic stability assessment, and antiviral evaluation of novel betulinic acid derivatives as potent anti-human

- immunodeficiency virus (HIV) agents. *Journal of Medicinal Chemistry*, 52(10), 3248–58.
- Qiu, J., Stevenson, S. H., O’Beirne, M. J., & Silverman, R. B. (1999). 2,6-Difluorophenol as a bioisostere of a carboxylic acid: bioisosteric analogues of gamma-aminobutyric acid. *Journal of Medicinal Chemistry*, 42(2), 329–32.
- R Development Core Team, R. (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (Vol. 1, p. 409).
- Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., ... Uetz, P. (2014). The binary protein-protein interaction landscape of Escherichia coli. *Nature Biotechnology*, 32(3), 285–90.
- Ramasubbu, N., Parthasarathy, R., & Murray-Rust, P. (1986). Angular preferences of intermolecular forces around halogen centers: preferred directions of approach of electrophiles and nucleophiles around carbon-halogen bond. *Journal of the American Chemical Society*, 108, 4308–4314.
- Ren, J., He, Y., Chen, W., Chen, T., Wang, G., Wang, Z., ... Xu, Y. (2014). Thermodynamic and structural characterization of halogen bonding in protein-ligand interactions: A case study of PDE5 and its inhibitors. *Journal of Medicinal Chemistry*, 57, 3588–3593.
- Reynolds, C. H., & Holloway, M. K. (2011). Thermodynamics of ligand binding and efficiency. *ACS Medicinal Chemistry Letters*, 2(6), 433–7.
- Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6), 276–277.
- Ross, G. A., Morris, G. M., & Biggin, P. C. (2012). Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites. *PLoS ONE*, 7(3), 1–13.
- Ross, N. T., Katt, W. P., & Hamilton, A. D. (2010). Synthetic mimetics of protein secondary structure domains. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 368(1914), 989–1008.
- Rye, C., & Baell, J. (2005). Phosphate Isosteres in Medicinal Chemistry. *Current Medicinal Chemistry*, 12(26), 3127–3141.
- Röckö, T., Tervo, A. J., Parkkinen, J., & Poso, A. (2006). BRUTUS: Optimization of a grid-based similarity function for rigid-body molecular superposition. II. Description and characterization. *Journal of Computer-Aided Molecular Design*, 20, 227–236.
- Şanlı, N., Şanlı, S., Özkan, G., & Denizli, A. (2010). Determination of pKa values of some sulfonamides by LC and LC-PDA methods in acetonitrile-water binary mixtures. *Journal of the Brazilian Chemical Society*, 21(10), 1952–1960.
- Satchell, J. F., & Smith, B. J. (2002). Calculation of aqueous dissociation constants of 1,2,4-triazole and tetrazole: A comparison of solvation modelsElectronic supplementary information (ESI) available: Calculated coordinates, atomic radii and charges of 1,2,4-triazole and tetrazole. See ht. *Physical Chemistry Chemical Physics*, 4(18), 4314–4318.
- Schottel, B. L., Chifotides, H. T., & Dunbar, K. R. (2008). Anion- π interactions. *Chemical Society Reviews*, 37(1), 68–83.
- Schwans, J. P., Sunden, F., Lassila, J. K., Gonzalez, A., Tsai, Y., & Herschlag, D. (2013). Use of anion-aromatic interactions to position the general base in the ketosteroid

- isomerase active site. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 11308–13.
- Seio, K., Miyashita, T., Sato, K., & Sekine, M. (2005). Synthesis and Properties of New Nucleotide Analogues Possessing Squaramide Moieties as New Phosphate Isosters. *European Journal of Organic Chemistry*, 2005(24), 5163–5170.
- Sengupta, D., Behera, R. N., Smith, J. C., & Ullmann, G. M. (2005). The alpha helix dipole: screened out? *Structure (London, England : 1993)*, 13(6), 849–55.
- Shapiro, C. L. (2009). Not boring at all. *EMBO Report*, 10(2), 125–128.
- Shen, Q., Xiong, B., Zheng, M., Luo, X., Luo, C., Liu, X., ... Jiang, H. (2011). Knowledge-based scoring functions in drug design: 2. Can the knowledge base be enriched? *Journal of Chemical Information and Modeling*, 51(2), 386–97.
- Sirivolu, V. R., Vernekar, S. K. V., Marchand, C., Naumova, A., Chergui, A., Renaud, A., ... Wang, Z. (2012). 5-Arylidenethioxothiazolidinones as inhibitors of tyrosyl-DNA phosphodiesterase I. *Journal of Medicinal Chemistry*, 55(20), 8671–84.
- Sisay, M. T., Peltason, L., & Bajorath, J. (2009). Structural interpretation of activity cliffs revealed by systematic analysis of structure-activity relationships in analog series. *Journal of Chemical Information and Modeling*, 49(10), 2179–89.
- Smith, F. W., Mudge, S. R., Rae, A. L., & Glassop, D. (2003). Phosphate transport in plants. *Plant and Soil*, 248(1/2), 71–83.
- Snyder, H. R., Reedy, A. J., & Lennarz, W. J. (1958). Synthesis of Aromatic Boronic Acids. Aldehydo Boronic Acids and a Boronic Acid Analog of Tyrosine 1. *Journal of the American Chemical Society*, 80(4), 835–838.
- Snyder, P. W., Mecinovic, J., Moustakas, D. T., Thomas, S. W., Harder, M., Mack, E. T., ... Whitesides, G. M. (2011). Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. *Proceedings of the National Academy of Sciences of the United States of America*, 108(44), 17889–94.
- Soldano, K. L., Jivan, A., Nicchitta, C. V., & Gewirth, D. T. (2003). Structure of the N-terminal domain of GRP94. Basis for ligand specificity and regulation. *The Journal of Biological Chemistry*, 278(48), 48330–8.
- Sparks, R. B., Polam, P., Zhu, W., Crawley, M. L., Takvorian, A., McLaughlin, E., ... Combs, A. P. (2007). Benzothiazole benzimidazole (S)-isothiazolidinone derivatives as protein tyrosine phosphatase-1B inhibitors. *Bioorganic and Medicinal Chemistry Letters*, 17, 736–740.
- Stumpfe, D., Dimova, D., & Bajorath, J. (2014). Composition and topology of activity cliff clusters formed by bioactive compounds. *Journal of Chemical Information and modelling*, 54(2), 451–461.
- Sun, H., Tawa, G., & Wallqvist, A. (2012). Classification of scaffold-hopping approaches. *Drug Discovery Today*, 17(7–8), 310–324.
- Sun, H., Zhao, L., Peng, S., & Huang, N. (2014). Incorporating replacement free energy of binding-site waters in molecular docking. *Proteins: Structure, Function and Bioinformatics*, 82(9), 1765–1776.
- Thornber, B. C. W. (1979). Isosterism and Molecular Modification in Drug Design. *Chemical Society Reviews*, 8, 563–580.

- Tian, L., Wu, Z., Zhao, Y., Meng, Y., Si, Y., Fu, X., ... Han, W. (2009). Differential induction of LRP16 by liganded and unliganded estrogen receptor alpha in SKOV3 ovarian carcinoma cells. *The Journal of Endocrinology*, 202(1), 167–77.
- Torricce, M. M., Bower, K. S., Lester, H. A., & Dougherty, D. A. (2009). Probing the role of the cation-pi interaction in the binding sites of GPCRs using unnatural amino acids. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 11919–11924.
- Tulsi, N. S., Downey, A. M., & Cairo, C. W. (2010). A protected 1-bromophosphonomethylphenylalanine amino acid derivative (BrPmp) for synthesis of irreversible protein tyrosine phosphatase inhibitors. *Bioorganic & Medicinal Chemistry*, 18(24), 8679–86.
- Ullmann, J. R. (1976). An Algorithm for Subgraph Isomorphism. *Journal of the ACM*.
- W.D. Kumler; John J. Eiler. (1943). The Acid Strength of Mono and Diesters of Phosphoric Acid. The *n*-Alkyl Esters. *Journal of the American Chemical Society*, 209(10), 2355–2361.
- Vagner, J., Qu, H., & Hruby, V. J. (2008). Peptidomimetics, a synthetic tool of drug discovery. *Current Opinion in Chemical Biology*, 12(3), 292–6.
- Vainio, M. J., & Johnson, M. S. (2007). Generating conformer ensembles using a multiobjective genetic algorithm. *Journal of Chemical Information and Modeling*, 47(6), 2462–74.
- Vainio, M. J., Puranen, J. S., & Johnson, M. S. (2009). ShaEP: molecular overlay based on shape and electrostatic potential. *Journal of Chemical Information and Modeling*, 49(2), 492–502.
- Van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2), 22–30.
- Wassermann, A. M., & Bajorath, J. (2011). Identification of target family directed bioisosteric replacements. *MedChemComm*, 2(7), 601–606.
- Waters, M. L. (2002). Aromatic interactions in model systems. *Current Opinion in Chemical Biology*, 6(6), 736–741.
- Wei, B. Q., Weaver, L. H., Ferrari, A. M., Matthews, B. W., & Shoichet, B. K. (2004). Testing a flexible-receptor docking algorithm in a model binding site. *Journal of Molecular Biology*, 337(5), 1161–82.
- Wels, B., Kruijtz, J. A. W., Garner, K., Nijenhuis, W. A. J., Gispen, W. H., Adan, R. A. H., & Liskamp, R. M. J. (2005). Synthesis of a novel potent cyclic peptide MC4-ligand by ring-closing metathesis. *Bioorganic & Medicinal Chemistry*, 13, 4221–4227.
- White, C. J., & Yudin, A. K. (2011). Contemporary strategies for peptide macrocyclization. *Nature Chemistry*, 3(7), 509–24.
- Whitesides, G. M., & Krishnamurthy, V. M. (2005). Designing ligands to bind proteins. *Quarterly Reviews of Biophysics*, 38, 385–395.
- Vieth, M., Siegel, M. G., Higgs, R. E., Watson, I. A., Robertson, D. H., Savin, K. A., ... Hipskind, P. A. (2004). Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs. *Journal of Medicinal Chemistry*, 47, 224–232.

- Willett, P. (2013). Combination of similarity rankings using data fusion. *Journal of Chemical Information and Modeling*, 53(1), 1–10.
- Wirth, M., Zoete, V., Michielin, O., & Sauer, W. H. B. (2013). SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic Acids Research*, 41(Database issue), D1137–43.
- Wolters, L. P., Schyman, P., Pavan, M. J., Jorgensen, W. L., Bickelhaupt, F. M., & Kozuch, S. (2014). The many faces of halogen bonding: a review of theoretical models and methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(6), 523–540.
- Xhaard, H., Backström, V., Denessiouk, K., & Johnson, M. S. (2008). Coordination of Na(+) by monoamine ligands in dopamine, norepinephrine, and serotonin transporters. *Journal of Chemical Information and Modeling*, 48(7), 1423–37.
- Yang, Y., Hu, B., & Lill, M. A. (2014). Analysis of Factors Influencing Hydration Site Prediction Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model.* 54(10), 2987–2995.
- Young, T., Abel, R., Kim, B., Berne, B. J., & Friesner, R. A. (2007). Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proceedings of the National Academy of Sciences of the United States of America*, 104(3), 808–13.
- Zacharias, N. (2002). Cation- π interactions in ligand recognition and catalysis. *Trends in Pharmacological Sciences*, 23(6), 281–287.
- Zhang, X., Zuo, Z., Tang, J., Wang, K., Wang, C., Chen, W., ... Zhou, H.-B. (2013). Design, synthesis and biological evaluation of novel estrogen-derived steroid metal complexes. *Bioorganic & Medicinal Chemistry Letters*, 23(13), 3793–7.
- Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7), 2302–9.
- Zhao, C., Lou, Z., Guo, Y., Ma, M., Chen, Y., Liang, S., ... Rao, Z. (2009). Nucleoside monophosphate complex structures of the endonuclease domain from the influenza virus polymerase PA subunit reveal the substrate binding site inside the catalytic center. *Journal of Virology*, 83(18), 9024–30.
- Zhao, P., Lu, Y., & Han, W. (2010). Clinicopathological significance and prognostic value of leukemia-related protein 16 expression in invasive ductal breast carcinoma. *Cancer Science*, 101(10), 2262–8.
- Zhao, R. Y., Erickson, H. K., Leece, B. a, Reid, E. E., Goldmacher, V. S., Lambert, J. M., & Chari, R. V. J. (2012). Synthesis and biological evaluation of antibody conjugates of phosphate prodrugs of cytotoxic DNA alkylators for the targeted treatment of cancer. *Journal of Medicinal Chemistry*, 55(2), 766–82.
- Zhu, X., Kim, J. L., Newcomb, J. R., Rose, P. E., Stover, D. R., Toledo, L. M., ... Morgenstern, K. A. (1999). Structural analysis of the lymphocyte-specific kinase Lck in complex with non-selective and Src family selective kinase inhibitors. *Structure*, 7(6), 651–661.