SAPIENZA

UNIVERSITÀ DI ROMA

Dipartimento di Chimica e Tecnologie del Farmaco
Ph.D. in Pharmaceutical Sciences
(XXIII cycle 2007 – 2010)

Rational Design of Novel Antiviral Compounds Through Computational Approaches

Tutor:
Chiar.mo Prof. Rino Ragno

Candidate:
Ira Musmuça

për Andrean

INDEX

1.   COMPUTATIONAL DRUG DISCOVERY

Computational methods have become standard in today's medicinal chemistry toolkit. Judging from the table of contents of medicinal chemistry journals or conference programs, there is a seamless integration of these disciplines into the early drug discovery process. While being confined at their beginning to the visualization of potential drug interactions in the targeted enzyme binding, today these techniques permeate all aspects of drug discovery and are indispensable prerequisites to support medicinal chemistry programs.[1]

Computational tools can offer significant advantages over the conventional ones, being orders of magnitude cheaper and faster or offering the possibility to predict molecular behaviours that cannot be elucidated in any other way. Moreover, recent developments in software and hardware make it possible to simulate more and more complex molecular environments, widening the applicability of *in silico* studies from the interactions of small molecules with key protein residues, to the simulation of the dynamic evolution of complex biological systems with atomic resolution.[2]

Nowadays, various computational disciplines are available and they can be broadly classified as structure-based or ligand-based approaches. The former is used in the presence of three-dimensional information of the target (generally derived from crystallographic or NMR studies). The latter methodologies are applied when the 3D structure of the target is not known, and information about the binding mode is derived from a series of related active compounds.

It is important to keep in mind that working with these techniques there is no single generic workflow, but is the combination of several methods that provide the best results. All methods should be seen as one multifaceted discipline contributing to the early drug

---

[1] Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813-1818.

[2] Zonta, N.; Coluccia, A.; Brancale, A. Advanced in silico approaches in antiviral research. *Antivir. Chem. Chemother.* **2010**, *20*, 147-151.

discovery process. However, the computer can just help in preparing the data, and at the end it is up to the project team to make use of all available information to ensure that the right choices are made and the most promising candidates are progressed.

1.1. Structure – Based Approaches

Structure-Based Drug Design (SBDD), is a well – established, successful and highly attractive strategy that uses three-dimensional protein structure information in the development of new biologically active molecules. As a creative and knowledge – driven approach, an essential requirement for structure-based studies is a substantial understanding of the spatial and energetic aspects that affect the binding affinities of protein – ligand complexes. Considering that the shape and the chemical nature of the binding site of a specific target protein are known, and the possible intermolecular interactions between ligands and the protein within its active site have been identified, this qualified information can be directly employed for the identification of new ligands and the optimization of lead compounds. This opens new possibilities to boost the search up for new lead molecules and to limit the number of compounds that need to be evaluated experimentally.[3,4]

Molecular docking and scoring technology is one of the most used computational technique when the three-dimensional structure of the desired target is available. In particular, docking can be defined as the positioning of a ligand in a protein binding site, and scoring, the quality assessment of docked ligands. This technology is applied at different stages of the drug discovery process for three main purposes: (1) predicting the binding mode of a known ligand; (2) identifying new ligands using virtual screening (VS); (3) predicting the binding affinities of related compounds from a known active site series. Of these three challenges, successful prediction of

---

[3] Andricopulo, A. D.; Salum, L. B.; Abraham, D. J. Structure-based drug design in medicinal chemistry. *Curr. Top. Med. Chem.* **2009**, *9*, 771-790.

[4] Cavasotto, C. N.; Orry, A. J. Ligand docking and structure-based virtual screening in drug discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1006-1014.

a ligand binding mode in a protein active site is perhaps the most straightforward and is the area where most success has been achieved.[5] Docking a ligand into a binding site models several degrees of freedom. These are the six degrees of translational and rotational freedom of one body relative to another and then the conformational degrees of freedom of the ligand and of the protein. The first docking algorithms only considered translational and orientation, with both ligand and protein treated as rigid bodies. Increases in computer performance and new algorithms enabled the ligand conformation degrees of freedom (ligand flexibility) to be explored.[5] Most docking programs today treat the ligand as flexible with a rigid (or nearly rigid) receptor structure. Receptor flexibility still remains one of the major challenges for the field.

Ligand Flexibility Algorithms – Usually, it is not known which conformation of a ligand interacts most favourably with a receptor and, therefore, a conformational searching during docking is necessary. Ligand flexibility algorithms can be divided in three types of searches, namely systematic, stochastic, and deterministic searches. Some algorithms use more than one of these approaches. Systematic search algorithms are based on a grid of values for each formal degree of freedom, and each of these grid values is explored in a combinatorial fashion during the search. An example of a systematic search is the anchor-and-grow or incremental construction algorithm.[6] Stochastic search algorithms make random changes usually changing one degree of freedom of the system at a time. One of the major concerns with stochastic searches is the uncertainty of convergence. In order to improve convergence, multiple, independent runs can be performed. Examples of stochastic searches are Monte Carlo (MC) methods and evolutionary algorithms. In deterministic searches the initial state determines the move that can be made to generate the next state, which generally has to be equal to or lower in energy than the initial

---

[5] Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851-5855.

[6] Broojimans, N.; Kuntz, I. D. Molecular recognitions and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335-373.

state. Deterministic searches performed on exactly the same starting system (including each degree of freedom) with the same parameters will generate exactly the same final state. A problem with deterministic algorithms is that they often get trapped in local minima because they cannot traverse barriers. Example of deterministic methods are energy minimization methods and molecular dynamics (MD) simulations.[6]

Receptor Flexibility Methods – Allowing ligand flexibility was a huge step forward, but changes in receptor structure upon ligand binding are observed frequently as well.[7] Actually, many molecular systems undergo rearrangements upon ligand binding, which range from local motions of side-chains to large domain movements. In any case, receptor flexibility might have a dramatic impact in the ligand docking problem as well as in virtual screening. Therefore, the inclusion of some structural variability in a docking study becomes of fundamental importance. However, the representation of protein flexibility is still a challenge for the state-of-the-art flexible ligand docking protocols.

Cross-docking is a feasible technique that accounts for protein flexibility and how it affects docking experiments per se.[8,9] Compared to self-docking studies, cross-docking analysis better verifies how efficiently docking studies can support lead identification and structure-activity relationship studies where no structural information are experimentally available for the ligands of interest. In this case, the ligand of interest is docked into an available protein structure that is usually complexed with a different ligand (e.g., a ligand A from a receptor-ligand A complex is docked into a receptor B from receptor-ligand B complex, where the receptor is the same biological target). Cross-docking analysis essentially simulates this situation because it evaluates how

[7] Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209-225.

[8] Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *J. Comput. Aided Mol. Des.* **2003**, **16**, 855-869.

[9] Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins* **1999**, *37*, 228-241.

efficiently a docking software package is able to reproduce the experimentally determined binding pose of a ligand by docking it into a protein whose 3D structure was determined in complex with a different ligand.[10]

Methods that specifically deal with receptor flexibility also exist. Some of them use conformational libraries to sample receptor flexibility.[11,12] Other methods use to pre-calculate receptor contributions to the binding free energy on a grid, a procedure that significantly increases the speed of docking algorithms. However, the grid also constricts docking to a rigid receptor because the grid is only calculated once. Different approaches try to overcome this problem by incorporating multiple receptor structures into the grid.[13,14,15] Some methods that sample ligand flexibility are also suited to sample receptor flexibility, namely Genetic Algorithms (GAs), MD methods, and MC methods.[6]

Scoring Functions – Once the configurations of a system are sampled, docking programs must evaluate and correctly rank the predicted ligand conformations. Even when binding conformations are correctly predicted, the calculations ultimately do not succeed if they do not differentiate correct poses from incorrect ones, and if 'true' ligands cannot be identified. So, the design of reliable scoring functions and schemes is of fundamental importance.

There are three important applications of scoring functions in molecular docking. The first of these is the determination of the binding mode and site of a ligand on a protein. Given a protein target, molecular docking generates hundreds of thousands of

[10] Tuccinardi, T.; Botta, M.; Giordano, M.; Martinelli, M. Protein Kinases: Docking and Homology Modeling Reliability *J. Chem. Inf. Model.* **2010**, *50*, 1432-1441.

[11] Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235*, 345-356.

[12] Jackson, R. M.; Gabb, H. A.; Sternberg, M. J. E. Rapid refinement of protein interfaces solvation: application to the docking problem. *J. Mol. Biol.* **1998**, *276*, 265-285.

[13] Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *76*, 269-281.

[14] Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002**, *46*, 34-40.

[15] Rarey, M.; Kramer, B.; Lengauer, T. Fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.

putative ligand binding orientations/conformations at the active site around the protein. A scoring function is used to rank these ligand orientations/conformations by evaluating the binding tightness of each of the putative complexes. An ideal scoring function would rank the experimentally determined binding mode most highly. Given the determined binding mode of a ligand, scientists would be able to gain a deep understanding of the molecular mechanism of ligand binding and to further design an efficient drug by modifying the protein or ligand.[16]

The second application of a scoring function, which is related to the first application, is to predict the absolute binding affinity between protein and ligand. This is particularly important in lead optimization.[17] Lead optimization refers to the process to improve the tightness of binding for low-affinity hits or lead compounds that have been identified. During this process, an accurate scoring function can greatly increase the optimization efficiency and save costs by computationally predicting the binding affinities between the protein and modified ligands before the much more expensive step of ligand synthesis and experimental testing.

The third application, perhaps the most important one in structure-based drug design, is to identify the potential drug hits/leads for a given protein target by searching a large ligand database, i.e., virtual database screening.[18] A reliable scoring function should be able to rank known binders most highly according to their binding scores during database screening. Given the expensive cost of experimental screening and sometimes unavailability of high-throughput assays, virtual database screening has played an increasingly important role in drug discovery.

All of these three applications, ligand binding mode identification, binding affinity prediction, and virtual database screening, are

[16] Rajamani, R.; Good, A. C. Ranking poses in structure-based lead discovery and optimization: Current trends in scoring function development. *Curr. Opin. Drug Discov. Dev.* **2007**, *10*, 308-315.

[17] Shoichet, B.K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead discovery using molecular docking *Curr. Opin. Chem. Biol.* **2002**, *6*, 439-446.

[18] Seifert, M. H.J.; Kraus, J.; Kramer, B. Virtual high-throughput screening of molecular databases. *Curr. Opin. Drug Discov. Dev.* **2007**, *10*, 298-307.

related to each other. Presumably, an accurate scoring function would perform equally well on each of them. Despite over a decade of development, scoring is still an open question. Many existing scoring functions perform well only on one or two of the three applications. Roughly, the scoring functions can be grouped into three basic types according to how they are derived: force field-based (i.e., a function expressing the energy of a system as a sum of diverse molecular mechanics or other terms), empirical, and knowledge-based.

Force field (FF) scoring functions are developed based on physical atomic interactions, including van der Waals interactions, electrostatic interactions, and bond stretching/bending/torsional forces. Force field functions and parameters are usually derived from both experimental data and *ab initio* quantum mechanical calculations according to the principles of physics. Despite its lucid physical meaning, a major challenge in the force field scoring functions is how to treat the solvent in ligand binding.[19]

Empirical scoring functions are fit to reproduce experimental data, such as binding energies and/or conformations, as a sum of several parameterized functions. The design of empirical scoring functions is based on the idea that binding energies can be approximated by a sum of individual uncorrelated terms. The coefficients of the various terms are obtained from regression analysis using experimentally determined binding energies and, potentially, X-ray structural information.[19]

A third kind of scoring functions are knowledge-based scoring functions (also referred to as statistical-potential based scoring functions), which employ energy potentials that are derived from the structural information embedded in experimentally determined atomic structures essentially designed to reproduce experimental

---

[19] Huang, Sh. Y.; Grinter, S. Z.; Zou, X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys*. **2010**, *12*, 12899-12908.

structures rather than binding energies.[20] So, in common with empirical methods, knowledge-based scoring functions attempt to implicitly capture binding effects that are difficult to model explicitly.[21] In Chart 1.1 a number of currently used scoring functions are summarized.

Force-field based	Empirical	Knowledge-based
D-score[9] G-score[9] Gold[22] AutoDock[23] DOCK[24]	LUDI[25] F-score[15] ChemScore[26] SCORE[27] Fresno[28] X-SCORE[29]	PMF[30] DrugScore[31] SMoG[32]

Chart 1.1 Types of scoring functions used in various docking software.

[20] Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **1990**, *213*, 859-883.

[21] Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935-949.

[22] Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J. Murray, C. W.; Taylor, R. D. Improved protein-ligand using GOLD. *Proteins* **2003**, *52*, 609-623.

[23] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical free energy scoring function. *J. Comput. Chem.* **1998**, *19*, 327-346.

[24] Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **2001**, *15*, 411-428.

[25] Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **1994**, *8*, 243-256.

[26] Eldrige, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V. Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **1997**, *11*, 425-445.

[27] Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: a new empirical method for estimating the binding affinity of a protein-ligand complex. *J. Mol. Model.* **1998**, *4*, 379-394.

[28] Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.* **1999**, *42*, 4650-4658.

[29] Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.* **2002**, *16*, 11-26.

[30] Muegge, I. A knowledge-based scoring function for protein-ligand interactions: probing the reference state. *Perspect. Drug Discov. Des.* **2000**, *20*, 99-114.

[31] Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337-356.

[32] DeWitte, R. S; Shakhnovic, E. I. SMoG: *de novo* design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733-11744.

Structure-based virtual screening – Docking and scoring is a technology that is also used for identifying new ligands using virtual screening (VS). This term, which was coined in the late 1990s, describes the use of computational algorithms and models for the identification of novel bioactive molecules.[33] VS methodology may involve searching databases containing hundreds-of-thousands, if not millions, of molecules and has rapidly become an essential component of the modern drug discovery process.[11,34,35,36,37,38]

Structure-based virtual screening (SBVS) approaches explore information about the target protein structure in order to select molecules that are likely to favourably interact and that can be further selected for *in vitro* biological tests. High-performance hardware and specialized software, combined with advanced knowledge of three-dimensional protein structure and small-molecule binding modes, have made this technology a useful complement, and in some cases a reasonable alternative to HTS.[4,34]

SBVS typically encompasses a sequence of crucial computational steps, including target preparation, chemical database selection, docking, scoring, post-docking analysis, ranking, visual inspection and priorization of compounds for testing as schematically shown in Figure 1.1. The knowledge-based SBVS approach is strongly affected by the quantity and quality of the information about the systems under investigation. The process of selection and preparation of the macromolecular target involves essential issues, such as druggability of the target receptor, selection of the most relevant geometry, flexibility, assignment of

[33] Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* **2010**, *9*, 273-276.

[34] Guido, R. V. C.; Oliva, G.; Andricopulo, A. D. Virtual screening and its integration with modern drug design technologies. *Curr. Med. Chem.* **2008**, *15*, 37-46.

[35] Joseph-McCarthy, D.; Baber, J. C.; Feyfant, E.; Thompson, D. C.; Humblet, C. Lead optimization via high-throughput molecular docking. *Curr. Opin. Drug Discov. Dev.* **2007**, *10*, 264-274.

[36] Lyne, P. D. Structure-based virtual screening: An overview. *Drug Discov. Today* **2002**, *7*, 1047-1055.

[37] Klebe, G. Virtual Ligand Screening: Strategies, perspectives and limitations. *Drug Discov. Today* **2006**, *11*, 580-594.

[38] Ghosh, S.; Nie, A.; An, J.; Huang, Z. Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol.* **2006**, *10*, 194-202.

charges, protonation and tautomeric states, ionization, and the inclusion of conserved water molecules in the binding cavity. Some relevant molecular characteristics such as partial charges, stereochemistry, ionization and tautomeric states must also be correctly assigned to the small molecule compounds.[21,34,36,39]
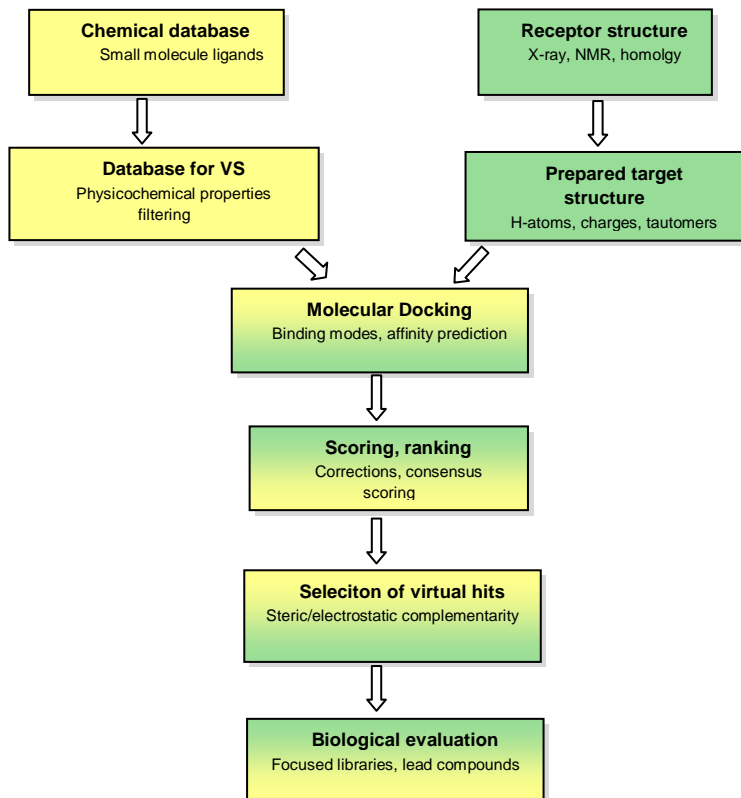

Figure 1.1. Flowchart of the computational steps in Structure-Based VS strategy.

Regardless the source of the small-molecule database, screening libraries generally contain a large number of molecules

---

[39] Klebe, G. Recent developments in Structure-Based Drug design. *J. Mol. Mod.* **2000**, *78*, 269-281.

with broad chemical diversity. In the important process of library design, several molecular filters can be used to reduce the number of compounds to be screened. Common filtering methods are variations of Lipinski's rule of five that include physicochemical and pharmacokinetic parameters in order to guide the selection of compounds with lead-like, fragment-like, and drug like properties. The chemical space can also be reduced by taking into account molecular properties presented by series of modulators with known biological activity, or through the identification of specific features required for ligand binding. Additional filters are often applied to remove specific chemical structures associated with chemical instability and toxicity.[36,40,41] All of these computational (virtual) screening filters are useful to improve the quality of small-molecule libraries and are crucial to the application of SBVS methods.

Receptor-based pharmacophore modelling – The term pharmacophore has been used in medicinal chemistry for many years. The official 1998 IUPAC definition of a pharmacophore is as follows: '*A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response*'.[42]

The most straightforward way to derive a 3D pharmacophore from a protein structure is through the direct observation of specific interactions between protein and ligand(s). Such a pharmacophore can then be used in the usual way, for example, to search a 3D database in order to identify compounds for focused screening. Database searching methods based on 3D pharmacophores are in general much faster than structure-based methods such as docking, and so this can be a more effective way to screen very large databases. Alternatively, the 3D pharmacophore search can act as

[40] Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening – An overview. *Drug Discov. Today* **1998**, *3*, 160-178.

[41] Walters, W. P.; Namchuk, M. Designing screens: How to make your hits a hit. *Nat. Rev. Drug Discov.* **2003**, *2*, 259-266.

[42] Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of terms used in medicinal chemistry (IUPAC recommendations). *Pure Appl. Chem.* **1998**, *70*, 1129-1143.

the first stage in a docking workflow. A large database can be pre-screened, using a pharmacophore query, to create a much smaller subset for docking. A pharmacophore, if properly constructed, will give very few false negatives (false negatives might arise from different binding modes or poor feature definition). The scores of docked hits may then become more useful.

Knowledge of the protein structure enables the pharmacophore to incorporate more detailed information about regions that are not accessible to the ligand. This is most commonly achieved through the use of exclusion volumes and/or inclusion regions.[43,44] When used for database searching, such pharmacophore queries have the distinct advantage of finding hits that not only contain the key binding elements but are also able to fit in to the active site, thereby reducing the false positive rate.

Structure-based 3D pharmacophores derived solely on the interactions observed in known protein-ligand complexes may be unnecessarily restrictive. An alternative is to define pharmacophores based on an analysis of the 'hot spots' in the active site. A number of methods can in principle be used to identify such hot spots (or site points). These include programs such as GRID[45] that probe the site with small molecules or functional groups and calculate the enthalpy of the interaction between the probe and the protein atoms at points on a grid lattice to generate a molecular interaction field (MIF). These fields can then be contoured by energy to find the most favourable regions for an acceptor or a donor (or any other type of feature) to interact with the protein. Programs such as LUDI[46] and

[43] Greenidge, P. A.; Carlsson, B.; Bladh, L.-G.; Gillner, M. Pharmacophores incorporating numerous excluded volumes defined by X-ray crystallographic structure in three-dimensional database searching: application to the thyroid hormone receptor. *J. Med. Chem.* **1998**, *41*, 2503-2512.

[44] Tintori, C.; Corradi, V.; Magnani, M.; Manetti, F.; Botta, M. Targets looking for drugs: a multistep computational protocol for the development of structure-based pharmacophores and their applications for hit discovery. J. *Chem. Inf. Model.* **2008**, *48*, 2166-2179.

[45] Goodford, P. J. A computational-procedure for determining energetically favourable binding-sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849-857.

[46] Bohm, H. J. The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J. Comput. Aided Mol. Des.* **1992**, *6*, 61-78.

SUPERSTAR[47] use a knowledge-based approach in which rules are used to generate a set of interaction sites for each atom or functional group of the protein that is capable of participating in a non bonded contact. The rules are largely based on statistical analysis of experimental structures from the Protein Data Bank or small molecule crystal structures and take into account the chemical nature of the atoms as well as the orientational preferences of features such as hydrogen bond donors/acceptors. From the locations of the site points it is then necessary to construct one or more 3D pharmacophores. The simplistic approach would be to combine all such points into a single pharmacophore, but the resulting query would typically be matched (if at all) only by a molecule filling the entire active site. Rather, all possible 3D pharmacophores (containing three, four, or more features) are enumerated from the set of site points. The most important step in such a procedure is to triage the set of pharmacophore queries. Commercially available programs that can perform the entire process from site to pharmacophores include Structure-Based Focusing[48] and LigandScout.[49]

***De novo* drug design** (DNDD) – DNDD is another interesting structure-based methodology. This approach can be described as the process of producing novel molecular structures with desired pharmacological properties, using the structural information of the biological target.[50] It is a fascinating field, but its use has been somehow restricted by some inherent limitations of the technique, in particular the inability of computers of assessing accurately the chemical feasibility of the proposed structures. For this reason, often an extensive user intervention in the design process is required.

---

[47] Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: a knowledge based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **1999**, *289*, 1093-1108.

[48] Venkatachalam, C. M.; Kirchoff, P.; Waldman, M. Receptor-Based Pharmacophore Perception and Modeling. In Pharmacophore Perception, Development and Use in Drug Design; Guner, O.F., Ed.; IUL Biotechnology Series: La Jolla, CA, 2000; pp 339-350.

[49] Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160-169.

[50] Schneider G, Fechner U. Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discov.* **2005**, *4*, 649-663.

1.2. Ligand – Based Approaches

Quantitative Structure – Activity Relationship (QSAR) predictions are possibly the most significant example of ligand-based drug design. These methods have been applied for decades in the development of relationships between physicochemical properties of chemical substances and their biological activities to obtain a reliable statistical model for prediction of the activities of new chemical entities. QSAR, in simple terms, is a method for building computational or mathematical models which attempts to find a statistically significant correlation between structure and function using a chemometric technique. In terms of drug design, structure here refers to the properties or descriptors of the molecules, their substituents or interaction energy fields, function corresponds to an experimental biological/biochemical endpoint like binding affinity, activity, toxicity or rate constants, while chemometric method include multiple linear regression (MLR),[51] partial least squares analysis (PLS),[52] principal component analysis (PCA),[53] principal component regression (PCR),[54] artificial neural networks (ANN),[55] genetic analysis (GA),[56] etc. Various QSAR approaches have been developed gradually over a time span of more than a hundred years and have served as a valuable predictive tool, particularly in the design of pharmaceuticals and agrochemicals. The methods have evolved from Hansch (1962) and Free-Wilson's one or two-dimensional linear free-energy relationships (1964), via Cramer's

[51] Berk, R. A. The formalities of multiple regression. In: *Regression Analysis: A Constructive Critique*; Berk, R. A. Ed.; SAGE Publications Ltd: London, 2003, pp. 103-110.

[52] Wold, S.; Johansson, E.; Cocchi, M. PLS: Partial least squares projections to latent structures. In: *3D QSARr in Drug Design: Theory, Methods and Applications*; Kubinyi, H. Ed.; ESCOM Science Publishers: Leiden, 1993, pp: 523-550.

[53] Dunteman, G. H. Basic Concepts of Principal Components Analysis. In: Principal Component Analysis; Dunteman, G. H. Ed.; SAGE Publications Ltd: London, 1998, pp: 15-22.

[54] Dunteman, G. H. Basic Concepts of Principal Components Analysis. In: Principal Component Analysis; Dunteman, G. H. Ed.; SAGE Publications Ltd: London, 1998, pp: 65-74.

[55] Baskin, II; Payulin, V. A.; Zefirov, N. S. Neural networks in building QSAR models. *Methods Mol. Biol.* **2008**, *458*, 137-158.

[56] Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structre-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854-866.

three-dimensional QSAR (1988) to Hopfinger's fourth[57] and Vedani's fifth and sixth-dimensions.[58,59] Irrespective of the type, all QSAR formalisms presume that every molecule included in the study binds to the same site of the same target receptor.[60]

Objectives of QSAR – Mostly all QSAR methods focus on the following goals:

- To quantitatively correlate and recapitulate the relationships between trends in chemical structure alterations and respective changes in biological endpoint for comprehending which chemical properties are most likely determinates for their biological activities.
- To optimize the existing leads so to improve their biological activities.
- To predict the biological activities of untested and sometimes yet unavailable compounds.[60]

Rationale behind QSAR Modelling – The extent of reliability in opting for QSAR modelling depends on the nature of property being predicted, the stage of the project and the relative ease and cost of compound synthesis and subsequent testing. Often QSAR models provide useful information but many times they fail, despite of good statistics generated from internal data used in training. Regardless of all such problems, QSAR becomes a useful alternative because of the following reasons:

- Conventional synthesis methods are expensive and time-consuming.

[57] Hopfinger, A. J.; Wang, Sh.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, J. P.; Duraiswami, Ch. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509-10524.

[58] Vedani, A.; Dobler, M. 5D-QSAR: The key for simulating induced fit? *J. Med. Chem.* **2002**, *45,* 2139-2149.

[59] Vedani, A.; Dobler, M.; Lill, M. A. Combining protein modeling and 6D-QSAR — Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.* **2005**, *48,* 3700-3703.

[60] Verma, J.: Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in Drug Design – A Review. *Curr. Top. Med. Chem.* **2010**, *10*, 95-115.

- Biological assays are often too costly, time and/or sacrifice of animals requiring, or compounds in their pure forms.
- Drug failures due to poor ADMET profiles at later stages of development (or even after commercialization) are exceedingly expensive and painful.
- Large number of compounds are available due to combinatorial chemistry and HTS approaches, but estimations for prioritization of synthesis and screening.[60]

   Classification of QSAR methodologies. *Based on dimensionality* – Frequently, QSAR methods are categorized into the following classes, based on the structural representation or by the way of derivation of the descriptor values:[60]

- 1D-QSAR, correlating biological activity with global molecular properties like pK_a, logP etc.
- 2D–QSAR, correlating biological activity with structural patterns like connectivity indices, 2D-pharmacophores etc., without taking into account the three dimensional representations of these properties.
- 3D-QSAR, correlating biological activity with non-covalent interaction fields surrounding the molecules.
- 4D-QSAR, additionally including ensemble of ligand configurations in 3D-QSAR.
- 5D-QSAR, that explicitly represent different induced-fit models in 4D-QSAR.
- 6D-QSAR, further incorporating different solvation models in 5D-QSAR.

Based on the type of chemometric method used – Sometimes QSAR methods are also classified into following two categories, depending upon the type of correlation technique employed to establish a relationship between structural properties and biological activity:

- Linear methods including linear regression (LR),[61] multiple linear regression (MLR),[51] partial least-squares (PLS),[52] and principal component analysis/regression (PCA/PCR).[53,54]
- Non-linear methods of artificial neural networks (ANN),[55] k-nearest neighbours (kNN),[62] and Bayesian neural nets.[63]

Limitations of classical QSAR methodologies – Classical QSAR methods are much simpler, faster and more amenable automation than 3D-QSAR approaches. They include clearly-defined physiochemical descriptors and are best suited for the analysis of large number of compounds and computational screening of molecular databases. Though they have been used for decades to correlate and predict the activity of molecules, they suffer from serious limitations in certain situations some of which are as follows:[64]

- Only 2D – structures are considered.
- Unavailability of appropriate physiochemical parameters (e.g., numerical descriptors for new or unusual substituents), rendering the compound unfit for inclusion in QSAR analysis.
- Insufficient parameters for describing drug-receptor interactions (e.g., steric parameter E_s, Hammett constant σ, etc.).
- Confined to only few substitutions in a common reference structure (simple variation of aromatic substituents) and works best with a congeneric series.
- No representations of stereochemistry or 3D-structure of molecules, regardless of their availability.
- Provide no unique solutions.

---

[61] Berk, R. A. Simple linear regression. In: *Regression Analysis: A constructive critique*. Berk, R. A., Ed.; SAGE Publications Ltd: London, 2003, pp: 21-38.

[62] Ajmani, S.; Jadhav, K.; Kulkarni, S. A. Three-dimensional QSAR using the k-nearest neighbour method and its interpretation. *J. Chem. Inf. Model.* **2006**, *46*, 24-31.

[63] Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42*, 3183-3187.

[64] Kubinyi, H. 2D QSAR models: Hansch and Free-Wilson Analyses. In: Comput. Med. Chem. Drug Discov. Bultinck, P.; Winter, H. D.; Langenaeker, W.; Tollenaere, J. P. Eds.; Marcel Dekker, Inc: New York, USA, 2004, pp. 539-570.

- Higher risk of chance correlations.
- High risk of failure due to 'too far outside' predictions.
- No graphical output thereby making the interpretation of results in familiar chemical terms, frequently difficult if not impossible.
- Requires considerable knowledge of substituent constants in physical organic chemistry to design a molecule, since classical QSAR equation do not directly suggest new compounds to synthesize.

Progress in 3D-QSAR approaches 3D-QSAR is a broad term encompassing all those methods which correlate macroscopic target properties with computed atom-based descriptors derived from the spatial (three-dimensional) representation of the molecular structures.[65,66,67] The methodology has emerged as a natural extension to the classical QSAR approaches pioneered by Hansch and Free-Wilson. The most frequently applied methods include comparative molecular field analysis (CoMFA),[68] comparative molecular similarity indices analysis (CoMSIA)[69] and the GRID/GOLPE program (generating optimal linear PLS estimations).[70,71]

[65] Akamatsu, M. Current state and perspectives of 3D-QSAR. *Curr. Top. Med. Chem.* **2002**, *2*, 1381-1394.

[66] Hopfinger, A. J.; Tokarski, J. S. Three-dimensional Quantitative Structure-Activity Relationship Analysis. In: *Practical Application of Computer-Aided Drug Design*; Charifson, P. S. Ed.; Marcel Dekker, Inc.: New York, USA, 1997; pp. 105-164.

[67] Oprea, T. I. 3D QSAR modeling in drug design. In: *Computational Medicinal Chemistry for drug discovery*; Bultinck, P.; Winter, H. D.; Langenaeker, W.; Tollenaere, J. P. Eds.; Marcel Dekker, Inc.: New York, USA, 2004, pp. 571-616.

[68] Cramer, R.D. III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA) 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967

[69] Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130-4146.

[70] Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct. Act. Relat.* **1993**, *12*, 9-20.

[71] Cruciani, G.; Watson, K. A. Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J. Med. Chem.* **1994**, *37*, 2589-2601.

The primary aim of 3D-QSAR methods is to understand structure–activity relationships and to employ this knowledge in order to propose novel compounds with enhanced activity and selectivity profile for a specific therapeutic target. 3D-QSAR methods, are nowadays used widely in drug design, since they are computationally not demanding and afford fast generation of QSARs from which the biological activity of newly synthesized molecules can be predicted. The basic assumption in 3D-QSAR is that a suitable sampling of the molecular interaction fields around a set of aligned molecules might provide all the information necessary for understanding their biological activities.[72] The suitable sampling is achieved by calculating interaction energies between each molecule and an appropriate probe at regularly spaced grid points surrounding the molecules. The resulting energies derived from simple potential functions can then be contoured to give a quantitative spatial description of molecular properties. If correlated with biological activity, 3D-fields can be generated, which describe the contribution of a region of interest surrounding the ligands to the target properties. However there is one main difficulty in the application of 3D-QSAR methods: for a correct model, a spatial orientation of the ligands towards one another has to be found, which is representative for the relative differences in the binding geometry at the protein binding site. The success of a molecular field analysis is therefore determined by the quality of the choice of the ligand superimposition.[73,74,75,76] In most cases, the first step in a 3D-QSAR study is the generation of a reliable pharmacophore model. Many alignment strategies have been reported and compared for this purpose (a detailed comparison of different

[72] Kubinyi, H. QSAR and 3D-QSAR in Drug Design. *Drug Discov. Tod.* **1997**, *2*, 457-467.

[73] Kim, K. Non-linear Dependencies in CoMFA. *J. Comput. Aided Mol. Des.* **1993**, 7, 71-82.

[74] Folkers, G.; Merz, A.; Rognan, D. CoMFA: Scope and Limitations. In: *3D-QSAR in Drug Design. Theory, Methods and Applications*, Kubinyi, H., Ed.; ESCOM Science Publishers B.V., Leiden, Netherlands, 1993,

[75] Klebe, G.; Abraham, U.; On the Prediction of Binding Properties of Drug Molecules by Comparative Molecular Field Analysis. *J. Med. Chem.* **1993**, *36*, 70-80.

[76] Sippl, W. Development of Biologically Active Molecules by Combining 3D QSAR and Structure-based Design Methods. *J. Comput. Aided Mol. Des.* **2002**, *16*, 825-830.

methods can be found in Holtje et al. work).[77] Depending on the molecular flexibility and the structural diversity of the investigated compounds this task of unique pharmacophore generation becomes less feasible. Despite the difficulties concerning the molecular alignment many successful 3D-QSAR studies applying the 3D-QSAR approaches have been reported in the last few years.[78,79,80,81] A major factor for the ongoing enthusiasm for CoMFA-related approaches comes from the proven ability of several of these methods to correctly estimate the biological activity of novel compounds. However, very often the predictive ability of QSAR models is only tested in retrospective studies rather than taking the ability to design and develop novel bioactive molecules. Despite the known limitations of 3D-QSAR, the possibility to predict biological data is gaining respect as scientists realize that we are far away from the hoped-for fast and accurate forecast of affinity from (the structure of a) protein–ligand complexes by free-energy perturbation or empirical scoring methods.[21,82,83]

Assumptions in 3D-QSAR methods
- There is an underlying relationship between molecular structure and biological activity.
- Receptor binding is directly proportional to the biological activity. Differential effects on second messengers or other

[77] Holtje, H. D.; Sippl. W.; Rognan, D.; Folkers, G. Molecular Modeling – Basic Principles and Applications, Wiley-VCH, Weinheim, 2003.

[78] Podlogar, B. L.; Ferguson, D. M. QSAR and CoMFA: A perspective on the practical application to drug discovery. *Drug Des. Discov.* **2000**, *1*, 4-12.

[79] Pastor, M.; Cruciani, G.; Watson, K. A Strategy for the Incorporation of Water Molecules Present in a Ligand Binding Site into a Three-Dimensional Quantitative Structure-Activity Relationship Analysis. *J. Med. Chem.* **1997**, *40*, 4089-4102.

[80] Sippl, W.; Contreras, J. M.; Parrot, I.; Rival, Y.; Wermuth, C.G. Structure based 3D-QSAR and Design of Novel Acetylcholinesterase Inhibitors. *J. Comput. Aided Mol. Des.* **2001**, *15*, 395-410.

[81] Audouze, K.; Ostergaard Nielsen, E.; Peters D. New Series of Morpholine and 1,4-Oxazepane Derivatives as Dopamine D4 Receptor Ligands: Synthesis and 3D-QSAR Model. *J. Med. Chem.* **2004**, *47*, 3089-3104.

[82] Masukawa, K. M.; Kollman, P. A.; Kuntz, I. D. Investigation of neuraminidase-substrate recognition using molecular dynamics and free energy calculations. *J. Med. Chem.* **2003**, *46*, 5628-5637.

[83] Tame, J. R. Scoring functions: The first 100 years. *J. Comput. Aided Mol. Des.* **2005**, *19*, 441-451.

signalling steps which transpire between receptor binding and experimentally observed response, are not taken into consideration.

- Molecular structure can be measured and represented with a set of numbers usually called descriptors, which encode all physical, chemical and biological properties of the molecule.

- Molecules with common or related structures generally have similar physicochemical properties (the similarity principle), and thus have similar binding modes and consequently comparable biological activities. The reverse also holds true. Similiarly, molecules located in the same region of the descriptor space present similar activity (the neighbourhood principle).

- Structural properties which lead to an observed biological response are most commonly determined by the non-bonding forces, mainly steric and electrostatic.

- The observed biological effect is produced by the modelled ligand itself, and not by its metabolite or degradation product.

- The lowest energy conformation of the ligand is its bioactive conformation, and it is this single conformation of the ligand which exerts the binding effects.

- With few exceptions, the geometry of the receptor binding site is considered rigid.

- The loss of translational and rotational degrees of freedom (entropy) upon binding is believed to follow a similar pattern for all the molecules.

- Total number of rotatable bonds is the only method most frequently used to estimate the entropic cost for freezing non-terminal single-bond rotors.

- For all modelled ligands, the protein binding site is assumed to be same.

- For all the modelled compounds, the on-off rate is supposed to be similar i.e., the system is considered to be in equilibrium, and kinetic aspects are usually ignored.

- Some of the major factors like desolvation energetic, temperature, diffusion, transport, pH, salt concentration etc. Which contribute to the overall free energy of binding are difficult to handle, and thus usually ignored.
- In molecular mechanics based 3D-QSAR methods, free energy of binding is largely explained by the enthalpic component (*i.e.*, the internal energy), which is prone to the inherent force field errors.
- Resulting QSAR model may represent one of potentially several solutions to the property-activity correlation problem.

Several parameters which significantly control the overall performance of a developed 3D-QSAR model are:

1. Biological data – The popular 'GIGO' (garbage in garbage out) principle applies in every computational technique. In 3D-QSAR also, one should utilize accurate activity data in order to develop a good model. Though 3D-QSAR methods can be applied to heterogeneous data sets, some considerations for maintaining the accuracy of biological data are necessary.[67,84]

1a. Compounds should belong to a congeneric series.

1b. Compounds should have the same mechanism of action and same/comparable binding mode.

1c. The biological activities of compounds should correlate to their binding affinity and their enumerated biological responses should be measurable.

1d. Biological data for molecules should be obtained using uniform protocols (radioligand, activator, cofactor, pH, buffer etc.) and preferably from a single source (organism/tissue/cell/protein) and single lab.

1e. Activity data for all the compounds should be in same units of measurement (binding/functional/

---

[84] Kim, K. H. Comparative molecular field analysis (CoMFA). In: Molecular similarity in drug design; Dean, P. M. Ed.; Blackie Academic & Professional: Glasgow, UK, 1995, pp. 291-331.

IC_{50}/K_i). K_i value is preferred instead of the IC_{50} data, since it is independent of the substrate concentration.

1f. The ranges of bioactivity covered should be as large as possible, keeping the mode of action identical. Preferably, activity range should be much larger than the standard deviations of the data; more than 3 log units with an even spread of data is preferred.

1g. If possible, biological data should be symmetrically distributed around their mean, and their precision should be evenly distributed over its range of variation. If not, such skewness can be removed by log transforming the data and expressing it as log (1/C), where C refers to the molar concentration of drug producing a standard response. It is noteworthy that free energy change is proportional to the inverse log of concentration of the compound.

2. **Compounds selection and series optimization** – One of the major applications of QSAR is to optimize the existing leads by structural modifications so as to improve their activity and reduce/eliminate side effects. However, there are many issues to be taken care of while selecting substituents for the modification of compounds; some of the most important ones are given below:[66,84]

2a. The compounds/substituents selected should be convincingly different from the existing ones, so as to minimize collinearity among the variables.

2b. The chosen compounds/substituents should have the properties which behave independent of each other, thereby maximizing dissimilarity and orthogonality.

2c. The selection should be done in such a manner so as to map the substituent (descriptor) space with minimum number of compounds.

2d. Synthetic accessibility/feasibility of the selected compounds should also be taken into consideration.

3. **Optimization of the 3D-structure of the molecules** – An important issue in 3D-QSAR is how to generate and represent molecular structure for analysis. The problem can be resolved both by experimental as well as computational techniques.[84] A large number of well-resolved experimentally crystal structures are available in databases like Cambridge Structural Database[85] and Protein Data Bank.[86] The crystal structures offer the advantage that some conformational information about the flexible molecule is included. However, molecular modelling methods are particularly useful for compounds that have not been made or cannot even exists under normal conditions. Computationally the 3D structures can be generated by three methods: (i) manually by sketching the structures interactively in a 3D-computer graphics interface, (ii) numerically by using mathematical techniques like distance geometry, quantum or molecular mechanics, and (iii) by automatic methods that are often used for building 3D-structure databases.

After the generation of the starting 3D-molecular structures, their geometries are refined by minimizing their conformational energies using theoretical calculation methods. Commonly used structure optimization techniques include:

- *Molecular Mechanics* methods which usually does not explicitly consider the electronic motion, and thus are fast, reasonably accurate and can be used for very large molecules like enzymes.

- *Quantum Mechanics* or *ab initio* methods which takes into account the 3D distribution of electrons around the nuclei, and therefore are extremely accurate but time consuming, computationally intensive and cannot handle large molecules.

- *Semi-empirical methods* which are basically quantum mechanical in nature but employs an extensive use of

---

[85] Allen, F. The Cambridge Structural Database: a quarter of a million crystal compounds and rising. *Acta Crystallogr. B* **2002**, *58*, 380-388.
[86] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Wissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res*. **2000**, *28*, 235-242.

approximation as in molecular mechanics. Generally, the molecular geometry is optimized by molecular mechanics methods, and its atomic charges are calculated mostly by semi-empirical or *ab initio* methods.

4. **Conformational analysis of molecules** – It is a well recognized fact that each compound containing one or more single bonds is existing at each moment in many different so-called rotamers or conformers. Although small molecules may have only a single lowest energy conformation, large and flexible molecules exist in multiple conformations at physiological conditions. Therefore, it becomes necessary to include various such conformations in a 3D-QSAR study.[84] Depending upon the type of molecules in the study, any of the following conformational search can be adopted:

    4a. Systematic search (or grid search) method which generates all possible conformations, by systematically varying each of the torsion angles of a molecule by some increment, keeping the bond lengths and bond angles fixed.

    4b. Random search method which generates a set of conformations by repetitively and randomly changing either the Cartesian (x, y, z) or the internal (bond lengths, bond angles and torsion/dihedral angles) coordinates of a starting geometry of the molecule under consideration.

    4c. Monte Carlo method which simulates dynamic behaviour of a molecule and generates the conformations by making random changes in its structure, calculating and comparing its energy with that of the previous conformation and accepting it if it is unique.

    4d. Molecular dynamics method which employs the Newton's second law of motion (force = mass x acceleration) to simulate the time-dependent movements and conformational changes in a

molecular system, and results in a so-called trajectory showing how the positions and velocities of atoms of atoms in the molecular system vary with time.

4e. Simulated annealing which heats up the molecular system under consideration to high temperature to overcome huge barriers, and after equilibrating there for some time using molecular dynamics, cools down the system slowly and gradually to obtain the low energy conformations according to the Boltzmann distribution.

4f. Distance geometry algorithm which generates a random set of coordinates by selecting random distances within each pair of upper and lower bounds to form constraints in a distance matrix, which are then used to generate energetically feasible conformations of a set of molecules.

4g. Genetic and evolutionary algorithms which are based on the concept of biological evolution and works by first creating a population of possible solutions to the problem. The solutions with best fitness scores crossovers and mutations over a time, and propagate their good characteristics down the generation to result in better solutions in the form of new conformers.

5. **Alignment of molecules** – One of the most crucial problems in most of the alignment-based 3D-QSAR methods is that their results are highly sensitive to the manner in which the bioactive conformations of all molecules are superimposed over each other.[66,84] In cases where all the molecules in a data set have a common rigid core structure, molecules can be aligned easily using least-square fitting procedure. However, in case of structural heterogeneity in the dataset, alignment of highly flexible molecules becomes quite difficult and time consuming. Several approaches have been

proposed to superimpose the molecules as accurately as possible, some of which are the following:

5a. Atom overlapping based superimposition: This method involves corresponding atom by atom pairing between molecules. It is also called the pharmacophore approach and is the most popular method since it gives the best matching of the preselected atom positions.

5b. Binding sites based superimposition: In this method, molecular alignment is obtained by superimposing the receptor active sites or the receptor residues that interact with the ligands.

5c. Fields/pseudofields based superimpositions: This method performs superimposition by comparing the similarities in the calculated interaction energy fields between the molecules.

5d. Pharmacophore based superimpositions: This method uses a hypothetical pharmacophore as a useful common target template. Each molecule is conformationally directed to assume the shape obligatory for its sub-molecular features to match with either a known pharmacophore or the one which is generated during the conformational analysis.

5e. Multiple conformers based superimposition: This method is particularly useful in cases where the ligands may bind to a receptor in multiple ways, or when the correct binding mode is unknown and the ligands have a fair degree of freedom of conformational flexibility.

Molecular Interaction Fields – Molecular interaction fields (MIFs) can be calculated for any molecule of known three-dimensional (3D) structure. A MIF describes the spatial variation of the interaction energy between a molecular target and a chosen probe. The target may be a macromolecule or a low molecular weight compound or a molecular complex. The probe may be a

molecule or a fragment of a molecule. MIFs can be applied in many ways.[87] They can guide the process of structure-based ligand design, may be used to dock ligands to macromolecules[88,89] and are also frequently used to derive quantitative structure–activity relationships (QSARs) for low molecular weight compounds.[68] Additionally, MIFs can be used to study the structure–activity relationships of macromolecules.[90,91] MIFs can also be applied to the prediction of pharmacokinetic properties, such as in the VolSurf methodology.[92]

In this work, MIFs will be described primarily with reference to their calculation with the GRID program.[93,94] Other programs may be used to compute MIFs; these have different energy functions and parametrizations.

Calculation of MIFs – **The target.** The starting point for a MIF calculation is provided by the atomic coordinates of the target molecule. These may have been determined experimentally or theoretically. In many calculations of MIFs, the target is treated as a rigid structure. However, in applications, it is often important to treat target flexibility, at least partially. There are several strategies for doing this:

[87] Wade, R. C. Molecular Interaction Fields, In: *3D QSAR in Drug Design. Theory, Methods and Applications*, Kubinyi, H. Ed.; ESCOM, Leiden, Netherlands, 1993, pp. 486-505.

[88] Wu, G.; Robertson, D. H.; Brooks, C. L. Detailed analysis of grid-based molecular docking: A case study of CDOCKERA CHARMm-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24*, 1549-1562.

[89] Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.* **1996**, *10*, 293-304.

[90] Schleinkofer, K.; Wiedemann, U.; Otte, L.; Wang, T.; Krause, G.; Oschkinat, H.; Wade, R. C. Comparative Structural and Energetic Analysis of WW Domain-Peptide Interactions. *J. Mol. Biol.* **2004**, *344*, 865-881.

[91] Wade, R. C.; Gabdoulline R. R.; De Rienzo, F. Protein Interaction Property Similarity Analysis. *Int. J. Quantum Chem.* **2001**, *83*, 122-127.

[92] Crivori, P.; Cruciani, G.; Carrupt, P. A.; Testa, B. Predicting blood-brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43*, 2204-2216.

[93] Goodford, P. J A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849-857.

[94] Goodford, P. J. GRID, Molecular Discovery, www.moldiscovery.com.

- Compute MIFs for multiple conformations of the target; the conformations may come from an NMR ensemble or from conformational searches or molecular dynamics simulations.
- Permit adaptation of the position of some atoms in the target to optimize the interaction energy of the probe during calculation of the MIFs. In the GRID program, this is routinely done for rotatable hydrogen atoms. It is also possible for the user to do calculations with specified side chains containing several non hydrogen atoms treated as movable in response to the probe position.

The target is also usually considered to have a single titration state and to be unaffected by the position of the probe. However, the GRID program does allow for probe-induced switching between histidine tautomers.

The target may consist of a single molecule or a complex of molecules or molecules and ions, such as e.g. for metalloproteins. Well ordered water molecules may also be considered part of the target. GRID also permits the possibility for a water-bridged target–probe interaction to be considered without a priori defining the position of the water molecule. The remaining solvent molecules are treated as a continuum that modulates the interaction energy between probe and target and may also have an entropic effect on the probe–target interaction.

The probe – MIFs are computed for positions of the probe at points on a rectilinear grid superimposed on the target. It is this grid that gives the GRID program its name. Grids of target–probe interaction energy values can be read into many molecular graphics programs which can display the MIFs as isoenergy contours or project the energies onto molecular surfaces. The probe is placed at each grid point in turn to compute a MIF. At its simplest, the probe is a unit positive charge representing a proton; in this case the MIF is the molecular electrostatic potential (MEP). Most probes are spheres parametrized to represent a specific atom or ion type. Hydrogen atoms are usually treated implicitly. For example, a carbonyl carbon atom and a methyl CH3 group are both treated as

spherical probes but the methyl probe has a larger radius because of the hydrogen atoms bonded to the carbon atom. Polar hydrogen atoms that can make hydrogen bonds can also be treated implicitly but the directional character of their hydrogen bonds must also be modelled. Nonspherical probes containing more than one non-hydrogen atom may also be used. In this case, the relation between the probe position and orientation and the grid point is not completely symmetric and has to be defined. In the GRID programme, a carboxylate group, for example, is treated as a 3-point probe with one of the oxygen atoms centred on the grid point. At each grid point, the probe is rotated around the grid point oxygen to energetically optimize its orientation.

The Interaction Function – Above, MIFs have been described as fields of probe–target interaction energies. The MIFs thus provide information on where the favourable and the unfavourable locations for the probe around the target are. The probe–target interaction function is typically an empirical molecular mechanics energy function. Its functional form is chosen to represent the underlying physical interactions and the function is parametrized against experimental observations. The energy function may contain explicit entropic terms or be parametrized such that it represents an interaction free energy rather than just an interaction energy. More recently, functions, sometimes referred to as potentials of mean force or knowledge-based functions, derived from statistical analysis of molecular structures have been derived to map out favourable and unfavourable locations for a probe with respect to a target molecule These functions, unlike an energy function, are not required to take a physically meaningful form. Given enough experimental data, they may reproduce experimental observations better than empirical energy functions but this can be at the disadvantage of a lack of analytical differentiability and of a transparent physical interpretation. They are most often used as scoring functions for ligand docking but some can be used to compute MIFs. The empirical energy functions used to compute

MIFs can consist of the sum of one or more terms. In the GRID program, the energy function is given by the following terms:

$$E = E_{VDW} + E_{EL} + E_{HB} + S$$

where EVDW is the van der Waals energy, EEL is the electrostatic energy, EHB is the hydrogen-bond energy and S is an entropy term.[95]

Data pretreatment and scaling – Usually in 3D-QSAR, before proceeding with the chemometric analysis, a pretreatment of the raw data is necessary in order to minimize redundancy.[84] One of the common reduction methods is based on the standard deviation cut-off, in which all the energy columns with a low standard deviation are eliminated from the data since they require longer computing time without contributing significantly to the results. Moreover, several variable selection methods are available and they can be used to reduce collinearity among the descriptors in order to prevent data over-fitting and to improve the prediction performance of the model.

After the pretreatment, the data is subjected to a scaling procedure which assigns equal weight to all the descriptors and places them on a common platform for a meaningful statistical analysis. Scaling significantly improves the signal to noise ratio and allows ranking of the individual variables based on their relative importance. Different scaling techniques are available and can be effectively used in 3D-QSAR approaches as for example:

Autoscaling, a method that scales the variables to a mean of zero and assigns them a unity variance (unit standard deviation) by dividing each column with its standard deviation;

Block-scaling, which provides each category of variables with the same weight by dividing initial autoscaled weights of descriptors in one class by the square root of the number of descriptors in that class;

---

[95] Cruciani, G. Methods and principles in medicinal chemistry – molecular interaction fields. VCH Publisher, New York, 2006.

Block-adjusted scaling, which is particularly useful when other variables are included along with the energy values in the analysis. This scaling gives other variables a comparable weight to the total variables.

Sometimes the pretreated data is subjected to centering by subtracting the column means from all the data. This does not change any coefficient values or comparative weights of the descriptors, but the number of significant components deriving from PLS may be smaller (one unit) than the number of significant components obtained from the data without centering. The method is supposed to improve the ease of interpretation and numerical stability.

Model Generation and Validation – After pretreatment and scaling, the molecular descriptors are then correlated to the biological activities of the molecules, assuming a liner relationship between them.[84,96] Since the number of independent (x) variables is much greater than the number of the compounds in the data set, the traditional linear regression analysis cannot be used to perform the fitting process. Therefore, to extract a stable and robust 3D-QSAR model from a range of possible solutions, the Partial Least Squares (PLS) technique is used.[52] PLS is an iterative regression procedure that produces its solutions based on linear transformation of a large number of original descriptors to a small number of new orthogonal terms called latent variables. PLS gives a statistically robust solution even when the independent variables are highly interrelated among themselves, or when the independent variables exceed the number of observations. Thus, PLS is able to analyze complex structure-activity data in a more realistic way, and effectively interpret the influence of molecular structure on biological activity. This is one of the standard statistical methods used for development of predictive 3D-QSAR models. Other methods to model linear relationships include MLR, PCA, PCR, etc.

---

[96] Cramer, R. D. III; Bunce, J. D.; Patterson, D. E.; Frank, I. E. Cross validation, bootstrapping and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct. Act. Relat.* **1998**, *7*, 18-25.

In spite of using PLS, spurious results can still occur due to the noise hidden in the obtained matrix. The GOLPE (which stands for generating optimal linear PLS estimation) approach was developed to identify which variables are meaningful for the prediction of the biological activity and to remove those with no predictivity.[70,71] Within this approach, fractional factorial design (FFD) is initially applied to test multiple combinations of variables.[71] For each combination, a PLS model is generated and only variables which significantly increase the predictivity are considered. Variables are then classified considering their contribution to predictivity. A further advance in GOLPE is the implementation of the smart region definition (SRD) procedure.[97] This is aimed at selecting the cluster of variables, rather than the single variable mainly responsible for activity. The SRD technique seems less prone to change correlation than any single variable selection, and improves the interpretability of the models.[79]

The most important criterion for judging the quality of a QSAR model is its ability to accurately predict not only the activities of the training set molecules (internal prediction), but also the activity of molecules that have been not included in the development of the statistical model (external prediction).[96] Various approaches used for this purpose are described:

- The correlation coefficient r, is a measure of the degree of linearity of the relationship. It signifies the quality of the models fitting and quantifies the variance in the data.[98] In an ideal situation the correlation coefficient must be equal to or approach 1, but actually due to the complexity of biological data, any value above 0.9 is appreciable. Correlation coefficients for the variables in a data set are compiled in a correlation matrix, which shows the relationship of one descriptor to another. The correlation matrix ensures that

[97] Pastor, M.; Cruciani, G.; Clementi, S. Smart Region Definition: A New Way to Improve the Predictive Ability and Interpretability of Three-Dimensional Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1997**, *40*, 1455-1464.

[98] Archdeacon, T. J. Regression and explained variance. In: *Correlation and Regression Analysis: a Historian's Guide*; Archdeacon, T. J.; Ed.; Univ. of Wisconsin Press: USA, 1994, pp. 178-196.

variables or significance are orthogonal to each other. The addition of every new variable to the model always increases the value of r, unless the new variable is a constant of a linear combination of other variables, which would not produce any effect. The increase in r caused by adding new variables signifies over-fitting of the data.

- The coefficient of multiple determinations also called Pearson correlation coefficient r^2, is the squared correlation coefficient which informs about how well the model reproduces the experimental data.98 It is a quantitative measure of the precision of adjustment for the fitted values and the observed ones. The closer it approaches to the unity, the bigger is the similarity of the adjusted values to the experimental ones, suggesting that the model fits the data unerringly. However, an r^2 close to 1 does not mean that the model is perfect; the addition of any new descriptor to the model induces an ever-increasing of r^2, even if the newly added descriptor does not contribute to the model. Thus, other measures are required to determine precisely the predictive capability of the model.

Cross–validation (CV) is one of the most extensively employed methods for the internal validation of a statistical model.[99] In cross – validation, the predictive ability of a model is estimated using a reduced data set of structural data. Usually, one element of the set is extracted each time, and a new model is derived based on the reduced data set, which is then employed to predict the activity of the excluded molecule. The procedure is repeated n number times until all compounds have been excluded and predicted once. This is the so called leave-one-out (LOO) method. Analogously, leaving out more than one molecule of the data set at a time is termed as leave-n-out or leave-many-out CV method.[96] The outcome of LOO

---

[99] Stone, M. Cross-validation choice and assessment of statistical predictions. *J. Stat. Soc. B.* **1974**, *36*, 111-147.

procedure is cross-validated correlation coefficient (r^2_{cv} or q^2) which is a criterion of both robustness and predictive ability of the model:

$$r^2_{cv} = (PRESS_0 - PRESS) / PRESS_0)$$

where $PRESS_0$ is the mean of the observed biological activity PRESS is the sum of the squares of the differences between the predicted and the observed activity values.[100] Many researchers consider high q^2 as the ultimate proof of high predictive power of the QSAR model which is generally considered as incorrect. Q^2 value should be regarded as a measure of internal consistency of the derived model rather than as a true indicator of the predictability. It should be noted that, since it is easier to fit the experimental data than to predict them from the QSAR model, r^2 of the model is always higher than q^2. Cross – validation is not foolproof. In highly redundant data sets with fewer degrees of freedom, it can give an over-optimistic result. It may also improperly indicate a lack of correlation if all the compounds in the data set are unique. Therefore, we can conclude that despite its wide acceptance, a high value of q^2 alone is an insufficient criterion for a QSAR model to be highly predictive.

Predictive ability of the model can also be evaluated by forecasting the activity of an external test set of molecules using the models derived from the training set. Predictive correlation coefficient (r^2_{pred}), which is analogous to cross-validated r^2 (or q^2) is a measure of the predictive ability of the derived QSAR model and is calculated by the following formula:[101]

$$r^2_{pred} = (SD - PRESS) / SD$$

where SD is the sum of squared deviations between the biological activities of the test set molecules and the mean activity of the training set molecules, while PRESS is the sum of squared deviations between the observed and the predicted activities of the test set molecules.

---

[100] Deep, R. Regression. In: *Probability and Statistics*; Deep, R., Ed.; Academic Press: UK, 2006, pp. 455-515.
[101] Marshall, G. R. Binding site modeling of unknown receptors. In: *3D QSAR in Drug Design: Theory, Methods and Applications*. Kubinyi, H.; Martin, Y. C.; Folkers, G. Ed.; Springer Publications, London, UK, 1998, pp: 80-116.

The **Fischer statistic** (F value) parameter is one of the several variance-related parameters that can be used as a measure of the level of statistical significance of the regression model.[102] A higher F value implies that a more significant correlation has been reached. It is used as a criterion to determine whether a more complex model is significantly better than a less complex one.

Display of results – One of the biggest advantages of 3D-QSAR over classical QSAR is the graphical interpretability of the statistical results. Equation coefficients can be visualized in the region around the ligands. Upon visual inspection, regions of space contributing most to the activity can be easily recognized. The interpretation of the graphical results allows one both to check the reliability of the models easily and to design modified compounds with improved activity or selectivity. In this respect 3D-QSAR methods like CoMFA and GRID/GOLPE have proven to be very useful.[72]

---

[102] Archdeacon, T. J. Evaluating the regression equation. In: *Correlation and Regression Analysis: a Historian's Guide*; Archdeacon, T. J.; Ed.; Univ. of Wisconsin Press: USA, 1994, pp. 178-196.

2.   ANTIVIRAL STRATEGIES: AN OVERVIEW

The study of viruses has historically provided and continues to provide the basis for much of our most fundamental understanding of modern biology, genetics, and medicine. Virology has had an impact on the study of biological macromolecules, processes of cellular gene expression, mechanisms for generating genetic diversity, processes involved in the control of cell growth and development as well as in the study of several aspects of molecular evolution.

In essence, viruses are collections of genetic information directed toward one aim: their own replication. They are the ultimate and prototypical example of 'selfish genes'. The viral genome contains the 'blueprints' for virus replication enciphered in the genetic code, and must be decoded by the molecular machinery of the cell that it infects to gain the objective. Viruses are, thus, obligate intracellular parasites dependent on the metabolic and genetic functions of living cells. This defining feature has a number of consequences for the development and application of antiviral drugs. The intracellular replication and the appropriation of cellular pathways for purposes of the pathogen makes it difficult to define virus-specific targets for therapeutic intervention, and inhibition strategies have to be highly specific to prevent cell toxicity. Furthermore, because of the viral dependence on suitable host cells and the fact that the pathogen is too small to be visible by light microscopy, complex systems are required for the propagation of viruses in the laboratory, the detection of virus replication, and the testing of potential inhibitors.

2.1. Principles of Viral Replication and Its Inhibition

Compared to bacteria and eukaryotic parasites, viruses are very simple pathogens. They have been described as 'a piece of bad news, wrapped in protein' – genomes encased by a protective shell composed of protein(s) and, in the case of enveloped viruses, of lipids. In contrast to mammalian cells or bacterial, fungal, or parasitic pathogens, viruses as a group do not share the same tupe of

genome or the principle of its replication. Viral genomes can consist of single- or double-stranded DNA or RNA, and viruses have been classified according to the type of genome and the genome replication strategy used (Fig. 2.1).
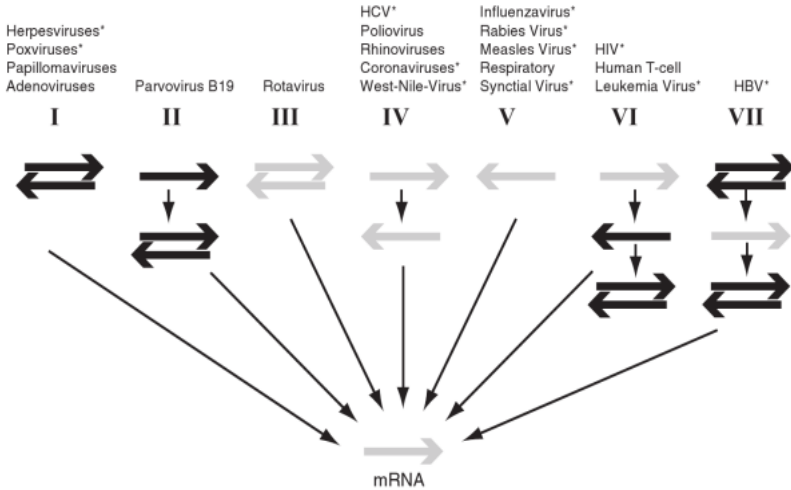

Figure 2.1. Classification of viruses by their genome replication strategy according to Baltimore. Examples for important human pathogens falling into the respective class are listed above. *Black*: DNA, *gray*: RNA; *arrows to the right*: (+) strand polarity (i.e., corresponding to mRNA); *arrows to the left*; (-)strands; *asterisk*: enveloped viruses.

Furthermore, they can be naked (i.e., containing only a protein shell) or enveloped by a lipid membrane that surrounds the protein shell and is derived from a host cell membrane.

Important human pathogens can be found in many different virus families. This has implications for antiviral intervention. While double-stranded DNA viruses largely use cellular pathways for genome replication, RNA viruses, or viruses replicating in the cytoplasm, have to provide own enzymes to mediate their virus-specific replication strategies. These enzymes represent targets for specific inhibition.

Although the details of the replication mechanism differ significantly between viruses, all viruses undergo the general replication steps outlined in Figure 2.2. First, the virion (defined as
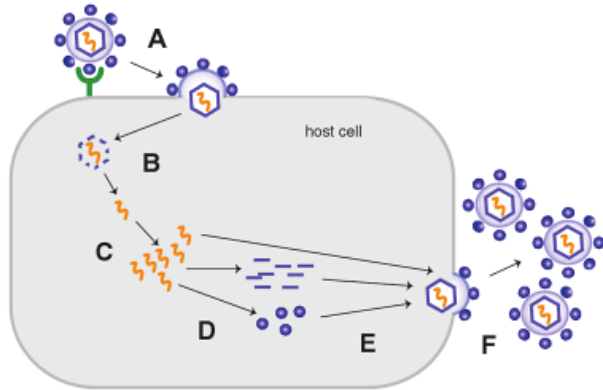

Figure 2.2. Basic steps of viral replication: (**a**) binding, (**b**) entry, (c) genome replication, (**d**) gene expression, (**e**) assembly, and (**f**) release.

the infectious viral particle) attaches to its host cell. Recognition of the appropriate target cell and binding is mediated by viral surface or envelope proteins interacting with one or more cellular membrane protein(s) and/or other attachment factor or plasma membrane (e.g., heparan sulfate proteoglycans, sialic acid). Subsequently, virus enter the cell either by uptake through a variety of endocytic pathways or directly at the plasma membrane. The following uncoating step, that is, the release of the viral genome from its protective proteinaceous shell, is currently poorly understood for most viruses. The subsequent replication of the viral genome occurs by different mechanisms for different viruses, depending on the type of the viral genome and the site of virus replication within the cell. Transcription, post-transcriptional modification of viral mRNA, and its translation are in most cases carried out by cellular machineries, but

these may be modifies by virus-encoded factors. Newly generated viral proteins and genomes are transported through cellular pathways to specific assembly sites within the host cell (e.g., the plasma membrane or intracellular virus factories), where they form progeny particles. Release of these particles can occur by cell lysis as observed for most naked viruses, or, in the case of enveloped viruses, by budding from a cellular membrane, leading to the acquisition of a lipid envelope.

In principle, each of these steps can be considered a target for antiviral intervention. Thus, inhibition of virus replication could be either accomplished by interfering with the specific virus-receptor interaction at the plasma membrane or the viral entry process, or by blocking viral enzymes involved in the genome replication or proteolytic maturation, or by affecting virus release. Targeting steps that rely mostly on cellular factors (transcription, translation, transport of viral components) is conceptually more difficult, while cellular factors or cellular machineries could present attractive targets due to their genetic stability if inhibited without major toxic effects. Specific interference with host cell factors or cellular machineries usurped by viruses may be achievable, provided that detailed knowledge on how the particular virus uses these pathways and on the viral and cellular factors involved is available. In addition to the direct interference with viral functions and factors, antiviral treatment can also involve immunomodulatory strategies.

2.2. Development of Antivirals

Several principles, requiring different levels of knowledge and methodology, can be applied to the identification of antiviral substances. First, a compound already in use as a therapeutic or known to inhibit another pathogen could be fortuitously discovered to inhibit the replication of a pathogenic virus. This principle is exemplified by the nucleoside analogs used as inhibitors of viral polymerases (see later) are related to antiproliferative drugs targeting cellular polymerases in highly replicating tumor cells. However, fortuitous discovery does obviously not represent an ideal

strategy for development of specific and affective drugs. In a more systematic approach, inhibitory compounds can be identified by random screening. For this purpose, a number of compound libraries compiled according to different principles are available from commercial and non-commercial sources. Random testing for inhibition of the replication of a given virus in tissue culture is rather unbiased towards specific inhibitory mechanisms and yields some preselection for compounds that do not display significant cytotoxicity and can be taken up into the cell. However, although tissue culture of many types of mammalian cells has become a routine method, screening in this setting is relatively expensive, time consuming, and not easily automated. Furthermore, not all viruses can be easily propagated in tissue culture: in the case of the important human pathogens hepatitis C virus (HCV) and hepatitis B virus (HBV), significant progress towards a routine tissue culture system has been accomplished only within last years, and only very recently have the systems been developed to allow medium to high throughput screening approaches.

Alternatively, a system for high-throughput random screening of antivirals can be set up by defining a specific viral target and establishing an *in vitro* assay appropriate to measure the function of this viral factor in the presence or absence of an inhibitors. Virus-encoded enzymes are particularly well suited targets for this approach, because in this case one can build on long-standing expertise from similar approaches in other areas of drug development, for example, metabolic diseases. However, *in vitro* systems mimicking non-enzyme mediated steps in virus replication, for example, virus assembly or protein-protein interactions between viral and cellular factors, are also being developed. Both types of random screening from compound libraries are not likely to identify a drug suited for treatment in the first round, but rather yield lead compounds that have to be validated by alternative assay procedures and subsequently improved in potency and pharmacological properties by iterative cycles of chemical modification and testing.

A fundamentally different approach is the procedure of rational drug design. Starting from detailed information on the molecular structure and function of a specific viral target, substances expected to bind to this target and to interfere with its function are identified by computer-aided drug design and subsequently synthesized and tested for inhibitory action. Again, suitable assay procedures to test for the inhibitory potential of *in silico* defined lead compounds have to be developed and the properties of the substance have to be improved by an iterative procedure. The application of *in-silico* technologies to antiviral research has already lead to the design of several compounds approved by the FDA and now used in clinical therapy.[103] Indeed, one of the earliest successes of these techniques is represented by the discovery of Zanamivir.[104] Palese *at al.* postulated that the sialidase enzyme was fundamental for influenza virus pathogenesis in man, although the first inhibitors of that enzyme failed to demonstrate their efficacy as anti-flu agents.[105] Only after the determination of the 3D structure of sialidase[105,106,107] the opportunity arose to design new selective inhibitors by structure-based approaches. Despite the initial limited commercial success of Zanamivir, the computer aided drug design strategies employed in its development were important first-steps in the development of further sialidase inhibitors.

Nowadays, many antiviral drug discovery programs include a molecular modelling component. The present work summarizes efforts made for the study and the development of new computational methodologies with the aim to discover new hits active versus hepatitis C virus (HCV).

---

[103] Congreve, M.; Murray, C. W.; Blundell, T. L. Structural biology and drug discovery. *Drug Discov. Tod.* **2005**, *10*, 895-907.

[104] Itzstein, V. M.; Wu, W.Y.; Kok, G. B. *et al.* Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363*, 418-423.

[105] Palese, P.; Schulman, J. L. In *Chemoprophylaxix and virus infection of the upper respiratory tract Vol. 1.* Ed.; JS Oxford. Cleveland: CRC Press; 1997, pp. 189-205.

[106] Varghese, J. N.; Colman, P. M. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature* **1983**, *303*, 35-40.

[107] Varghese, J. N.; Colman, P.M. Structural basis of antigenic variation: Studies of influenza virus neuraminidase. *J. Molec. Biol.* **1991**, *221*, 473-486.

3. HEPATITIS C VIRUS

Hepatitis C virus (HCV), the agent responsible for most cases of blood-borne hepatitis, was discovered by Choo *et al.* 20 years ago.[108] HCV is the leading cause of chronic liver disease worldwide. It is estimated that about 170 million people are chronically infected with HCV, and HCV accounts for approximately 20% of cases of acute hepatitis and 70% of cases of chronic hepatitis.[109] Chronic hepatitis C is a major cause of cirrhosis and hepatocellular carcinoma. Moreover, HCV-related end-stage liver disease is, in many countries, the first cause of liver transplantation.

The mechanisms responsible for the persistence of viral infection and for the liver lesions are not well understood. Despite an active immune response of the host, HCV has the capability to escape. It is believed that the quasispecies nature of HCV is one of the major mechanisms allowing the virus to cause chronic infection. HCV is not directly cytopathic and liver lesions are mainly related to immune mediated mechanisms. Co-factors influencing the outcome of the disease, including age, gender and alcohol consumption, are poorly understood and other factors, *e.g.*, immunologic and genetic, may play an important role.

Since the first study by Hoofnagle *et al.* in 1986,[110] many controlled trials have shown the efficacy of alpha interferon in patients with chronic hepatitis C with a sustained response in a minority of patients. Earlier studies have shown that the efficacy of treatment was significantly improved by the combination of interferon and ribavirin, which is nowadays the preferred treatment. Predictors of sustained response to therapy, mainly viral load and genotype, have been identified. Despite this remarkable progress

---

[108] Choo, Q.L.; Kuo, G.; Weiner, A. J.; Obverby, L. R.; Bradley, D. W.; Houghton, M. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science* **1989**, *244*, 359-362.

[109] Marcellin, P. Hepatitis C: clinical spectrum of the disease. *J. Hepatology* **1999**, *31*, 9-16.

[110] Hoofnagle, J. H.; Mullen, K. D.; Jones, D. B.; Rustgi, V.; Di Bisceglie, A.; Peters, M.; Waggoner, J. G.; Park, Y.; Jones, A. Treatment of chronic non-A, non-B hepatitis with recombinant human alpha interferon: a preliminary report. *N. Engl. J. Med.* **1986**, *315*, 1575-1578.

and the forthcoming use of improved forms of ribavirin,[111] it is unlikely that current treatment will succeed in curing all patients. Actually, besides the incomplete efficacy of genotypes 1,[112] combination therapy has significant side effects and is poorly tolerated in particular by individuals affected by other diseases. Consequently, many patients cannot be treated with current combination therapy and overall chances of cure for patients are probably below 50%. Given the high prevalence of the disease, the development of more effective, convenient and tolerated treatments is a major public health objective. Similar to the HIV antiviral research, efforts to develop antiviral agents for HCV have mostly focused on inhibition of key viral enzymes. With the aim of understanding the virology and pathogenesis of HCV, a description of the viral genome structure and replication will be discussed in this chapter.

3.1. Virology and Pathogenesis

Hepatitis C is caused by a small RNA virus that is included in the *Flaviviridae family* and has been classified as the sole member of the genes hepacivirus.[113] HCV has a 9.6 single-stranded RNA which encodes a single polyprotein of about 3000 aminoacids.[114] This HCV polyprotein is cleaved into a number of structural and non-structural proteins including: 2 envelope protein (core-C) and several non-structural (from NS2 to NS5) proteins.

Recent studies have classified some of the functions of these viral proteins. NS3 has helicase and protease activities while NS5 contains the RNA dependent RNA polymerase activity. All these enzymatic activities are essential for HCV replication and are

---

[111] Watson, J. Prospects for hepatitis C virus therapeutics: levovirin and viramidine as improved derivation of ribavirin. *Curr. Opin. Invest. Drugs*. **2002**, *3*, 680-683.

[112] Chander, G.; Sulkowski, M. S.; Jenckes, M. W.; Torbenson, M. S.; Bass, H. F. Treatment of chronic hepatitis C: a systematic review. *Hepatology* **2002**, *36*, S135-S144.

[113] Robertson, B.; Myers, G.; Howard, C.; Brettin, T.; Bukh, J.; Gaschen, B. Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. *Arch. Virol.* **1998**, *143*, 2493-2503.

[114] Major, M.; Feinstone, S. M. The molecular biology of hepatitis C. *Hepatology* **1997**, *25*, 1527-1538.

currently considered as targets for the development of new antiviral compounds.[115]

Some of the viral proteins have been recently implicated also in the pathogenesis of the liver disease and in the development of resistance to interferon therapy. The HCV core proteins, which exist as full-length and truncated forms, have been shown to regulate apoptosis of infected cells[116] and might be therefore directly implicated in the pathogenesis of liver disease, of cell proliferation and liver cancer development. The core proteins, as well as NS5A, have also been reported to interfere with intracellular metabolism of lipoproteins with a direct effect on the development of steatosis[117] which is a characteristic feature of hepatitis C. Last but not least, NS5A may contain a sequence domain able to regulate the cellular response to interferon. This region, which has been termed as interferon sensitivity determining region (ISDR) encodes for a NS5A protein sequence that can bind and inhibit protein kinase R (PKR), a protein whose activity is of fundamental importance in the development of an intracellular antiviral state in response to interferon.[118] Despite these advances in the understanding of HCV pathogenesis, the knowledge on HCV replication and viral proteins has been limited by the lack of suitable cell culture systems for expression and propagation of HCV and of simple small animal models of HCV infection.

Analysis and modelling of HCV kinetics during the early phase of interferon therapy seem to indicate that the rate of virus production is quite high in patients with hepatitis C, in the range of 10^{10} - 10^{12} virions per day, with a very rapid viral turnover and a predicted half

[115] Lauer, G. M.; Walker, B. D. Hepatitis C virus infection. *N. Engl. J. Med.* **2001**, *345*, 41-52.

[116] Ruggieri, A.; Harada, T.; Matsuura, Y.; Miyamura, T. Sensitization to Fas-mediated apoptosis by hepatitis C virus core protein. *Virolgy* **1997**, *229*, 68-76.

[117] Perlemuter, G.; Sabile, A.; Letteron, P.; Vona, G.; Toplico, A.; Chretien, Y.; Koike, K.; Pessayre, D.; Chapman, J.; Barba, G.; Brèchot, C. Hepatitis C virus core protein inhibits microsomal triglyceride transfer activity and very low density lipoprotein secretion: a model of viral-related steatosis. *FASEB J.* **2002**, *16*, 185-194.

[118] Tan, S. L.; Katze, M. G. How hepatitis C virus counteracts the interferon response: the jury is still out on NS5A. *Virology* **2001**, *284*, 1-12.

life of 2 -3 h.[119] Recently, the development of subgenomic replicons of HCV has provided a new research tool for hepatitis C. These *in vitro* systems can support efficient HCV-RNA replication and synthesis of all viral proteins, although unfortunately complete virus particles production has not yet been achieved. Currently, the replicon system is actively used for testing the antiviral effect of new antiviral drugs as all viral enzymes that are the major target foe antiviral therapy (NS2-3 and NS3 proteinases, NS3 helicase, NS5B RNA-dependent RNA-polymerase) are encoded by the replicon RNAs. On the other hand, in the absence of complete HCV particles production, these systems have limited impact in the understanding of HCV replication and pathogenesis.

Studies conducted during natural infection in humans indicate that chronicity of hepatitis C is related to rapid production of virus, and a lack of vigorous T-cell immune response to HCV with emergence of HCV variants which are prone to escape immune control.[120]

The pathogenesis of liver damage is most likely due to a combination of direct cytopathic effects of viral proteins and of immune mediated mechanisms including cytolytic and non-cytolytic reactions mediated by CTLs (cytotoxic T lymphocytes) and inflammatory cytokines.[121]

Recent data indicate that oxidative stress is an important pathogenic factor in HCV related liver damage.[122] Hepatitis steatosis is also a characteristic feature of hepatitis C and contributes to the progression of liver disease and fibrosis development.[123]

[119] Neumann, A. U.; Lam, N. P.; Dahari, H.; Gretch, D. R.; Wiley, T. E.; Layden, T. J.; Perelson, A. S. Hepatitis C. Viral dynamics *in vivo* and the antiviral efficacy of interferon-α therapy. *Science* **1998** *282*, 103-107.

[120] Farci, P.; Shimoda, A.; Coiana, A.; Diaz, G; Peddis, G.; Melpoldeer, J. C.; Strazzera, A.;. Chien, D. Y.; Munoz, S. J.; Balestrieri, A.; Purcell, R. H.; Alter, H. J. The outcome of acute Hepatitis C predicted by the evolution of the viral quasispecies. *Science* **2000**, *288*, 339-344.

[121] Cerny, A.; Chisari, F. V. Pathogenesis of chronic hepatitis C: immunological features of hepatic injury and viral persistence. *Hepatology* **1999**, *30*, 595-601.

[122] Okuda, M.; Li, K.; Beard, M. R.; Showalter, L. A.; Scholle, F.; Lemon, S. M.; Weinman, S. A. Mitochondrial injury, oxidative stress, and antioxidant gene expression are induced by hepatitis C virus core protein. *Gastroenterology* **2002**, *122*, 366-375.

[123] Monto, A.; Alonzo, J.; Watson, J. J.; Grunfeld, G.; Wright, T. L. Steatosis in chronic hepatitis C: relative contribution of obesity, diabetes mellitus and alcohol. *Hepatology* **2002**, *36*, 729-736.

3.2. Structural Biology of Hepatitis C Virus

Structural analyses of HCV components provide an essential framework for understanding of the molecular mechanisms of HCV polyprotein processing, RNA replication, and virion assembly and may contribute to a better understanding of the pathogenesis of hepatitis C. As abovementioned, studies of this virus have been hampered by the lack of a productive cell culture system; most information thus has been obtained from analysis of the HCV genome, heterologous expression systems, *in vitro* and *in vivo* models, and structural analyses.

The HCV genome is a single-stranded RNA molecule of positive polarity. It contains a single open reading frame (ORF) encoding a polyprotein of about 3,000 amino acids (Figure 3.1). The ORF is flanked by 5' and 3' untranslated regions (UTR) of 341 and approximately 230 nucleotides in length, respectively. Both 5' and 3' UTR bear highly conserved RNA structures essential for polyprotein translation and genome replication.[124] The structural proteins include the core (C), which forms the viral nucleocapsid, and the envelope glycoproteins E1 and E2. They are released by host-cell signals peptidases.

Most of the C protein is found in the cytoplasm, where it is bound to ER membranes or located at the surface of lipid droplets.[125] A small proportion of the core protein also may be found in the nucleus. In addition to its role in nucleocapsid formation, HCV core protein may modulate gene transcription, cell proliferation, cell death, and cell signalling; may interfere with lipid metabolism; and may suppress host immune responses.[126]

The two envelope glycoproteins, E1 and E2, are thought to play pivotal roles at different steps of HCV replication cycle. There is now strong evidence that they are essential for host-cell entry, by binding

---

[124] Penin, F.; Dubuisson, J.; Rey, F. A.; Moradpour, D.; Pawlotsky, J. M. Structural biology of Hepatitis C Virus. *Hepatology* **2004**, *39*, 5-19.

[125] Hope, R. G.; McLauchan, J. Sequence motifs required for lipid droplet association and protein stability are unique to the hepatitis C virus core protein. *J. Gen. Virol.* **2000**, *81*, 1913-1925.

[126] McLauchan, J. Properties of the hepatitis C virus core protein: a structural protein that modulates cellular processes. *J. Viral Hepat.* **2000**, *7*, 2-14.

to receptor(s) and inducing fusion with host-cell membrane.[127] As in other *Flaviviridae*, they are thought to play a major role in viral particle assembly.[128]
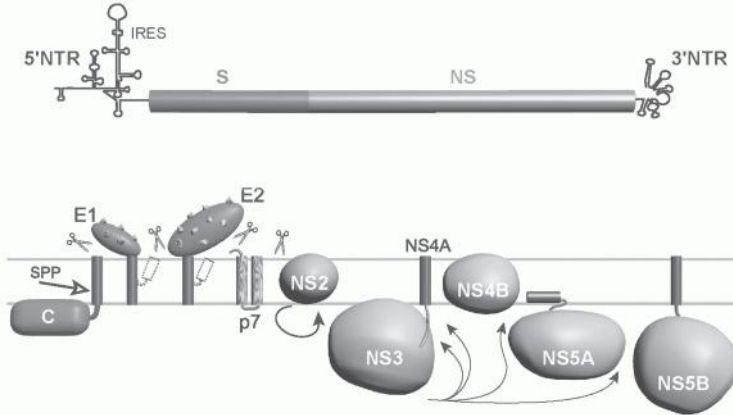

Figure 3.1. HCV genome organization (top) and polyprotein processing (bottom). The 5' UTR region consists of four highly structured domains and contains the internal ribosome entry site. The 3' UTR consists of stable stem-loop structures and an internal poly(U)/polypyrimidine tract. The central 9.6-kb ORF codes for a polyprotein of slightly more than 3000 aa depending on the HCV genotype. S and NS correspond to regions coding for structural and nonstructural proteins, respectively. The polyprotein processing and the location of the 10 HCV proteins relative to the ER membrane are schematically represented. Scissors indicate ER signal peptidase cleavage sites; cyclic arrow, autocatalytic cleavage of the NS2-NS3 junction; black arrows, NS3/NS4A proteinase complex cleavage sites; intramembrane arrow, cleavage by the signal peptide peptidase. The transmembrane domains of E1 and E2 are shown after signal-peptidase cleavage and reorientation of the respective C-terminus hydrophobic stretches (dotted rectangles).

These distinct functions imply that the envelope proteins adopt markedly different conformations and that the later must be

---

[127] Bartosch, B.; Dubuisson, J.; Cosset, F-L. Infectious hepatitis C pseudoviruses containing functional E1E2 envelope protein complexes. *J. Exp. Med.* **2003**, *197*, 633-642.

[128] Kuhn, R. J.; Zhang, W.; Rossman, M. G.; Pletnev, S. V.; Corver, J.; Lenches, E.; Jones, C. T., et al. Structure of dengue virus : implications for flavivirus organization, maturation, and fusion. *Cell* **2002**, *108*, 717-725.

controlled tightly to occur at the appropriate phases of the replicative cycle.

The nonstructural (NS) proteins NS2 to NS5B are involved in polyprotein processing and viral replication. The proteolytic processing of NS polyprotein part is complex and requires two distinct proteinases: the NS2–NS3 zinc–dependent metalloproteinase, and the NS3 serine protease located in the N-terminal region of NS3. The NS2–NS3 proteinase appears to be dedicated solely to cleavage at the NS2/NS3 site that occurs rapidly and by a conformation-dependent, autocatalytic mechanism.[129] The remaining NS proteins are released by the NS3 proteinase associated with its cofactor, NS4A. The C-terminal region of NS3 protein includes RNA helicase and NTPase activities.The NS3 serine proteinase domain associates with the NS4A cofactor, stabilizing the proteinase and activating it to cleave sites 4A/4B, 4B/5A/, and 5A/5B. The serine proteinase activity of NS3 is an attractive target for new drugs that block viral replication efficently. NS4B is an integral membrane protein of unknown function. NS5A is a membrane-associated polyphosphorylated protein of unknown function in the HCV replication complex that has attracted considerable interest because of its potential role in modulating the response to interferon alpha therapy.[130,131] Japanese studies suggested the existence of a discrete region of NS5A, tentatively named the *interferon sensitivity determining region*, and NS5A has also been reported to interfere with the activity of the double stranded-RNA-activated protein kinase *in vitro*.[132,133] However, these

[129] Hijikata, M.; Mizushima, H.; Akagi, T.; Mori, S.; Kakiuchi, N.; Kato, N.; Tanaka, T.; et al. Two distinct proteinase activities required for the processing of a putative non-structural precursor protein of hepatitis C virus. *J. Virol.* **1993**, *67*, 4665-4675.

[130] Pawlotsky, J. M. Hepatitis C virus resistance to antiviral therapy. *Hepatology*, **2000**, *32*, 889-896.

[131] Katze, M. G.; He, Y.; Gale, M. Jr. Viruses and interferon; a fight for supremacy. *Nat. Rev. Immunol.* **2002**, *2*, 675-687.

[132] Enomoto, N.; Sakuma, I.; Asahina, Y.; Kurosaki, M.; Murakami, T.; Yamamoto, C, Ogura, Y. et al. Mutations in the nonstructural protein 5° gene amd response to interferon in patients with chronic hepatitis C virus 1b infection. *N. Engl. J. Med.* **1996**, *334*, 77-81.

[133] Gale, M. Jr.; Blakely, C. M.; Kwieciszewski, B.; Tan, S. L.; Dosser, M.; Tang, N. M.; Korth, M. J. et al. Control of PKR protein kinase by hepatitis C virus non-structural 5A protein: molecular mechanisms of kinase regulation. *Mol. Cell Biol.* **1998**, *18*, 5208-5218.

findings are controversial, and the role of NS5A in the viral replication resistance to interferon-α remains elusive. Many other potential functions recently have been attributed to NS5A, including transcriptional activation and involvement in the regulation of cell growth and cellular signalling pathways.[134] Recently,[135] researchers at the Bristol-Myers Squibb R&D have reported two new iminothiazolidinone based compounds (**BMS-858** and **BMS-790052**) characterized with a potent HCV replication inhibitory activity. The antiviral activity of **BMS-790052** maps to the first 100 amino acids of HCV NS5A. These results provide clinical validation for the first in a new class of HCV inhibitors that target a viral protein with no known enzymatic function and an as yet poorly understood role in viral replication.

[134] Tellinghuisen, T. L.; Rice, C. M. Interaction between hepatitis C virus proteins and host cell factors. *Curr. Opin. Microbiol.* **2002**, *5*, 419-427.

[135] Gao, M.; Nettles, R. E.; Belema, M.; Snyder, L. B.; Nguyen, V. N.; Fridell, R. A.; Serrano-Wu, M. H.; Langley, D. R.; Sun, J-H.; O'Boyle II, D. R et al. Chemical genetics strategy identifies an HCV NS5A inhibitor with a potent clinical effect. *Nature* **2010**, *465*, 42-44.

4.   SMALL MOLECULE INTERFERON INDUCERS

Hepatitis C is becoming an increasingly common cause of mortality especially in the HIV-coinfected group. Due to the efficacy of interferon (IFN) based therapy in the treatment of Hepatitis C, various compounds possessing IFN-inducing activity have been hitherto reported.

In the present study,[136] we describe how steric, electrostatic, hydrophobic and hydrogen-bonding interactions might influence the biological activity of a published set of IFN inducers, using a Three-Dimensional Quantitative Structure–Activity Relationships (3D-QSAR) approach. Analyses were conducted evaluating different series of compounds structurally related to 8-hydroxyadenines and 1*H*-Imidazo[4,5-c]quinolines. A ligand-based alignment protocol in combination with the GRID/GOLPE approach was applied; 62 3-D QSAR models were derived using different GRID probes and several training sets. Performed 3D-QSAR investigations proved to be of good statistical value displaying r^2, $q^2_{CV\text{-}LOO}$ and cross-validated SDEP values of 0.73, 0.61, 0.61 and 0.89, 0.64, 0.58 using the OH or the DRY probe, respectively. Additionally, the predictive performance was evaluated using an external test set of 20 compounds. Analyses of the resulting models led to the definition of a pharmacophore model that can be of interest to explain the observed affinities of known compounds as well as to design novel low molecular weight IFN-inducers (IFNIs).

To the best of our knowledge, this is the first 3D-QSAR application on IFN-inducing agents.

4.1. Introduction

All current treatment protocols for hepatitis C are based upon interferon alpha (IFN-a) alone or in combination with Ribavirin.[137,138]

---

136 Musmuca, I.; Simeoni, S.; Caroli, A.; Ragno, R. Small-Molecule Interferon Inducers. Toward the Comprehension of the Molecular Determinants through Ligand-Based Approaches. *J. Chem. Inf. Model.* **2009**, *49*, 1777-1786

[137] Koike, K. Antiviral treatment of hepatitis C: present status and future prospects. *J. Infect. Chemother.* **2006**, *12*, 227-232.

Despite progress of therapeutic strategies such as the development of pegylated interferon,[139] sustained viral response (SVR) rates are still typically poor for genotype-1 infections that predominate in Europe, Japan and US too.[140] Additionally, the aforementioned therapy is often accompanied by significant problems:[138] i) recombinant IFNs, in fact, are genetically engineered bringing to the recognition by the body's immune system as "foreign". Consequently, the benefit of the treatment is strongly affected by the immune responses and the formation of neutralizing antibodies; ii) the cost of treatment is very high because of the inordinately expensive IFN preparations; iii) administration of IFN by intramuscular or subcutaneous injection frequently causes pain and irritation at the site of injection. Due to the above mentioned issues, the enhancement of the endogenous IFN release by the administration of orally available small molecular weight compounds is a promising approach in the anti-HCV therapy. Compounds possessing IFN-inducing activity have been already reported. Among them, small molecule IFN inducers (IFNIs) include tilorone,[141,142] BL-20803,[143] atabrine,[144] CP-28888,[145] ABMP,[146]

[138] Chevaliez, S.; Pawlotsky, J. M. Interferon-based therapy of hepatitis C. *Adv. Drug Deliv. Rev.* **2007**, *59*, 1222-1241.

[139] Hayashi, N.; Takehara, T. Antiviral therapy for chronic hepatitis C: past, present, and future. *J. Gastroenterol.* **2006**, *41*, 17-27.

[140] Dymock, B. W. Emerging therapies for hepatitis C virus infection. *Expert. Opin. Emerg. Drugs* **2001**, *6*, 13-42.

[141] Stringfellow, D. A.; Glasgow, L. A. Tilorone hydrochloride: an oral interferon-inducing agent. *Antimicrob. Agents Chemother.* **1972**, *2*, 73-78.

[142] Mayer, G. D.; Krueger, R. F. Tilorone hydrochloride: mode of action. *Science* **1970**, *169*, 1214-1215.

[143] Siminoff, P.; Bernard, A. M.; Hursky, V. S.; Price, K. E. BL-20803, a new, low-molecular-weight interferon inducer. *Antimicrob. Agents Chemother.* **1973**, *3*, 742-743.

[144] Glaz, E. T.; Szolgay, E.; Stoger, I.; Talas, M. Antiviral activity and induction of interferon-like substance by quinacrine and acranil. *Antimicrob. Agents Chemother.* **1973**, *3*, 537-541.

[145] Hoffman, W. W.; Korst, J. J.; Niblack, J. F.; Cronin, T. H. N,N-dioctadecyl-N',N'-bis(2-hydroxyethyl) propanediamine: antiviral activity and interferon stimulation in mice. *Antimicrob. Agents Chemother.* **1973**, *3*, 498-502.

[146] Nichol, F. R.; Weed, S. D.; Underwood, G. E. Stimulation of murine interferon by a substituted pyrimidine. *Antimicrob. Agents Chemother.* **1976**, *9*, 433-439.

DRB,[147] 10-carboxy- methyl-9-acridone,[148] bropirimine[149] and imiquimod (Chart 4.1).[150,151]
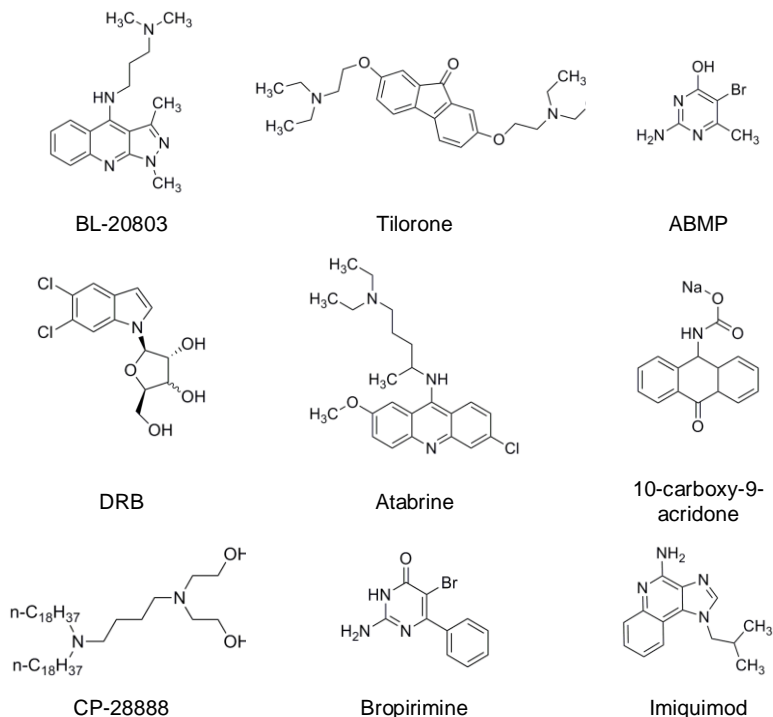

Chart 4.1. Chemical structures of small molecular weight IFNIs.

[147] Tamm, I.; Sehgal, P. B. A comparative study of the effects of certain halogenated benzimidazole ribosides on RNA synthesis, cell proliferation, and interferon production. *J. Exp. Med.* **1977**, *145*, 344-356.

[148] Taylor, J. L.; Schoenherr, C. K.; Grossberg, S. E. High-yield interferon induction by 10-carboxymethyl-9-acridanone in mice and hamsters. *Antimicrob. Agents Chemother.* **1980**, *18*, 20-26.

[149] Wierenga, W.; Skulnick, H. I.; Stringfellow, D. A.; Weed, S. D.; Renis, H. E.; Eidson, E. E. 5-substituted 2-amino-6-phenyl-4(3H)-pyrimidinones. Antiviral- and interferon-inducing agents. *J. Med. Chem.* **1980**, *23*, 237-239.

[150] Hirota, K.; Kazaoka, K.; Niimoto, I.; Kumihara, H.; Sajiki, H.; Isobe, Y.; Takaku, H.; Tobe, M.; Ogita, H.; Ogino, T.; Ichii, S.; Kurimoto, A.; Kawakami, H. Discovery of 8-hydroxyadenines as a novel type of interferon inducer. *J. Med. Chem.* **2002**, *45*, 5419-5422.

[151] Wierenga, W.; Skulnick, H. I.; Stringfellow, D. A.; Weed, S. D.; Renis, H. E.; Eidson, E. E. 5-substituted 2-amino-6-phenyl-4(3H)-pyrimidinones. Antiviral- and interferon-inducing agents. *J. Med. Chem.* **1980**, *23*, 237-239.

Imiquimod (**1**) is a well known IFN inducer (IFNI), clinically used in US for treatment of exophytic warts caused by the human papillomavirus.[150] As reported by Akira and co-workers,[152] Imiquimod acts via Toll-like receptor 7 (TLR7) signaling pathway inducing the endogenous IFN release. However, due to the serious adverse effects arose during clinical trials, Imiquimod based strategy against HCV[153] was abandoned. Even if, analogues were discovered to be more effective than Imiquimod, up to now no IFN inducers are clinically employed in the anti-hepatitis C therapy.[154] Three-Dimensional Quantitative Structure-Activity Relationships (3D-QSAR) continues to be a vigorous field and an increasingly attractive discipline in the scientific community. 3D-QSAR methods such as Comparative Molecular Field Analysis (CoMFA) and the GRID/GOLPE combination, have been successfully applied in many instances to guide the design of bioactive molecules. Although several series of IFNI*s* have been disclosed, to the best of our knowledge, no 3D-QSAR studies have been performed on these classes of compounds. The present study is the application of the GRID/GOLPE procedure to different series of IFNIs. Robust and predictive 3D-QSAR models were built in order to gain insights into how steric, electrostatic, hydrophobic, and hydrogen-bonding interactions might influence their biological activity. The aim of this work was the development of useful models for future designing and forecasting the activity of untested IFNI*s*.

4.2. Results

176 IFNIs structure-activity data were considered in the present

---

[152] Hemmi, H.; Kaisho, T.; Takeuchi, O.; Sato, S.; Sanjo, H.; Hoshino, K.; Horiuchi, T.; Tomizawa, H.; Takeda, K.; Akira, S. Small anti-viral compounds activate immune cells via the TLR7 MyD88-dependent signaling pathway. *Nat. Immunol*. **2002**, *3*, 196-200.

[153] Strominger, N. L.; Brady, R.; Gullikson, G.; Carpenter, D. O. Imiquimod-elicited emesis is mediated by the area postrema, but not by direct neuronal activation. *Brain Res. Bull*. **2001**, *55*, 445-451.

[154] Kurimoto, A.; Tobe, M.; Ogita, H.; Ogino, T.; Takaku, H.; Ichii, S.; Kawakami, H.; Isobe, Y. Prodrugs of 9-benzyl-8-hydroxy-2-(2-hydroxyethylthio)adenine: potent interferon inducing agents in monkeys. *Chem. Pharm. Bull. (Tokyo)* **2004**, *52*, 466-469.

study. The whole set of adenines[150,155,156,157,158] and 1H-imidazo[4,5-c]quinolines related structures (Figure 4.1) were selected from literature. Compounds and biological activities are detailed in Tables S1-S6 in the Supporting Information.
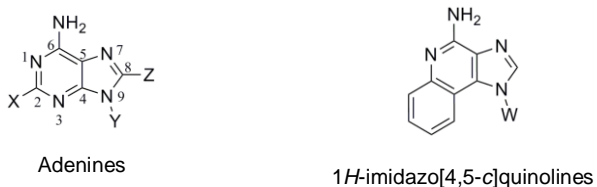

Adenines

1*H*-imidazo[4,5-*c*]quinolines

Figure 4.1. Adenine and 1*H*-imidazo[4,5-*c*]quinoline molecular scaffolds

IFNIs of Tables S1-S6 were firstly collected into five distinct groups according to both molecular scaffolds (1H-imidazo[4,5-c]quinolines and Adenines) and substituents at positions 2, 8 and 9 on the adenine ring, providing: 1H-imidazo[4,5-c]quinolines (group G1), 2-hydrogen/ alkyl/ cycloalkyl/ aryl-8-hydroxy/ mercapto-9-(substituted)arylalkyl adenines (group G2), 2-(alkyl)amino-8-hydroxy-9-benzyladenines (group G3), 2-alkylamino-8-hydroxy-9-(substituted)arylalkyl/ alkyl/ alkylhydroxy (group G4) and 2-alkoxy/ alkylthio-8-hydroxy-9-benzyladenines (group G5). G1-G5 groups chemical structures and biological activities are reported in Table 4.1 – 4.5, respectively. Activity data correspond to the minimum effective micromolar concentration required to observe IFN induction, as

---

[155] Isobe, Y.; Kurimoto, A.; Tobe, M.; Hashimoto, K.; Nakamura, T.; Norimura, K.; Ogita, H.; Takaku, H. Synthesis and biological evaluation of novel 9-substituted-8-hydroxadenine derivatives as potent interferon inducers. *J. Med. Chem.* **2006**, *49*, 2088-2095.

[156] Isobe, Y.; Tobe, M.; Ogita, H.; Kurimoto, A.; Ogino, T.; Kawakami, H.; Takaku, H.; Sajiki, H.; Hirota, K.; Hayashi, H. Synthesis and structure-activity relationships of 2-substituted-8-hydroxyadenine derivatives as orally available interferon inducers without emetic side effects. *Bioorg. Med. Chem.* **2003**, *11*, 3641-3647.

[157] Kurimoto, A.; Ogino, T.; Ichii, S.; Isobe, Y.; Tobe, M.; Ogita, H.; Takaku, H.; Sajiki, H.; Hirota, K.; Kawakami, H. Synthesis and structure-activity relationships of 2-amino-8-hydroxyadenines as orally active interferon inducing agents. *Bioorg. Med. Chem.* **2003**, *11*, 5501-5508.

[158] Kurimoto, A.; Ogino, T.; Ichii, S.; Isobe, Y.; Tobe, M.; Ogita, H.; Takaku, H.; Sajiki, H.; Hirota, K.; Kawakami, H. Synthesis and evaluation of 2-substituted 8-hydroxyadenines as potent interferon inducers with improved oral bioavailabilities. *Bioorg. Med. Chem.* **2004**, *12*, 1091-1099.

reported in the original papers. All derivatives[159] with activity values reported in µg/mL were transformed in µM to have uniformity of data.
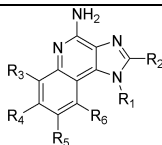
Compounds were then submitted to 3D-QSAR analyses with the aim of identifying and understanding structural properties and features able to influence their biological activity. In this context, starting with two dimensional text formats as input and using the stand-alone version of PRODRG program[160,161] in conjunction with the molecular mechanics of GROMACS,[162] coordinates and molecular topologies of the ligands were generated.

[159] Gerster, J. F.; Lindstrom, K. J.; Miller, R. L.; Tomai, M. A.; Birmachu, W.; Bomersine, S. N.; Gibson, S. J.; Imbertson, L. M.; Jacobson, J. R.; Knafla, R. T.; Maye, P. V.; Nikolaides, N.; Oneyemi, F. Y.; Parkhurst, G. J.; Pecore, S. E.; Reiter, M. J.; Scribner, L. S.; Testerman, T. L.; Thompson, N. J.; Wagner, T. L.; Weeks, C. E.; Andre, J. D.; Lagain, D.; Bastard, Y.; Lupu, M. Synthesis and structure-activity-relationships of 1H-imidazo[4,5-c]quinolines that induce interferon production. *J. Med. Chem.* **2005**, *48*, 3481-3491.
[160] Schuttelkopf, A. W.; van Aalten, D. M. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 1355-1363.
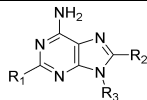[161] van Aalten, D. M.; Bywater, R.; Findlay, J. B.; Hendlich, M.; Hooft, R. W.; Vriend, G. PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *J. Comput. Aided Mol. Des.* **1996**, *10*, 255-262.
[162] Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43-56.

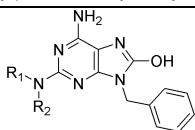**Table 4.1.** Group G1: 1H-imidazo[4,5-c]quinolines.

entry	R_1	R_2	R_3	R_4	R_5	R_6	MEC[a]
1	$CH_2CH(CH_3)_2$	H	H	H	H	H	2.08
2[c]	$CH_2CH(OH)CH_2OH$	H	H	H	H	H	9.68
3	CH_3	H	H	H	H	H	2.52
4	CH_2CH_3	H	H	H	H	H	2.36
5	$CH_2CH_2CH_3$	H	H	H	H	H	2.21
6	$CH(CH_3)_2$	H	H	H	H	H	2.21
7	$CH_2CH_2CH_2CH_3$	H	H	H	H	H	2.08
8	$CH_2(CH_2)_3CH_3$	H	H	H	H	H	1.97
9	$CH_2(CH_2)_4CH_3$	H	H	H	H	H	1.86
10	$CH_2(CH_2)_5CH3$	H	H	H	H	H	1.77
11	CH_2CH_2OH	H	H	H	H	H	2.19
12[c]	$CH_2CH(CH_3)OH$	H	H	H	H	H	1.03
13	$CH_2C(CH_3)_2OH$	H	H	H	H	H	1.95
14	CH_2Ph	H	H	H	H	H	0.36
15	CH_2CH_2Ph	H	H	H	H	H	1.73
16	$CH_2CH_2CH_2CH_2Ph$	H	H	H	H	H	0.32
17[b]	H	H	H	H	H	H	2.71
18	$CH_2CH(CH_3)_2$	H	H	H	H	H	0.2
19	$CH_2CH(CH_3)_2$	CH_2CH_3	H	H	H	H	0.19
20	$CH_2CH(CH_3)_2$	CH_2CH2CH_3	H	H	H	H	0.18
21b	$CH_2CH(CH_3)_2$	$CH_2CH2CH_2CH_3$	H	H	H	H	0.03
22	$CH_2CH(CH_3)_2$	$CH_2(CH_2)_3CH_3$	H	H	H	H	0.16
23	$CH_2CH(CH_3)_2$	$CH_2(CH_2)_4CH_3$	H	H	H	H	1.54
24	$CH_2CH(CH_3)_2$	$CH_2(CH_2)_5CH_3$	H	H	H	H	2.95
25[b]	$CH_2CH(CH_3)_2$	CH_2OPh	H	H	H	H	0.29
26[b]	$CH_2CH(CH_3)_2$	CH_2Ph	H	H	H	H	0.15
27	$CH_2CH(CH_3)_2$	CH_2CH_2Ph	H	H	H	H	2.9
28	$CH_2CH(CH_3)_2OH$	CH_3	H	H	H	H	0.18
29	$CH_2CH(CH_3)_2OH$	CH_2CH_3	H	H	H	H	0.18
30	$CH_2CH(CH_3)_2OH$	$CH_2CH(CH_3)_2$	H	H	H	H	0.03
31	Ph	$CH_2CH_2CH_2CH_3$	H	H	H	H	1.58
32	$CH_2CH(CH_3)_2$	H	OH	H	H	H	3.9
33	$CH_2CH(CH_3)_2$	H	H	CH_3	H	H	1.97
34	$CH_2CH(CH_3)_2$	H	H	OCH_3	H	H	3.7
35	$CH_2CH(CH_3)_2$	H	H	OH	H	H	1.95

[a] minimum effective concentration (μM); [b] compounds used in the test set; [c] compounds containing a stereogenic centre.

**Table 4.2.** Group G2: Adenine derivatives substituted at positions C-2, C-8 and N-9.

entry	R$_1$	R$_2$	R$_3$	MECa
36	H	OH	CH$_2$-Ph	10.0
37	H	SH	CH$_2$-Ph	10.0
38	H	OH	α-naphthylphenyl	10.0
39	H	OH	CH$_2$C$_6$H$_4$(4-F)	10.0
40	H	OH	CH$_2$C$_6$H$_4$(4-Cl)	10.0
41	H	OH	CH$_2$C$_6$H$_4$(4-CH$_3$)	10.0
42	H	OH	CH$_2$C$_6$H$_4$(2-OCH$_3$)	10.0
43	H	OH	CH$_2$C$_6$H$_4$(3-OCH$_3$)	10.0
44	H	OH	CH$_2$C$_6$H$_4$(4-OCH$_3$)	10.0
45	CH$_3$	OH	CH$_2$-Ph	1.0
46	C$_2$H$_5$	OH	CH$_2$-Ph	0.3
47	C$_3$H$_7$	OH	CH$_2$-Ph	0.1
48[b]	C$_4$H$_9$	OH	CH$_2$-Ph	0.01
49	Ph	OH	CH$_2$-Ph	10.0
50	C$_5$H$_{11}$	OH	CH$_2$-Ph	0.03
51	C$_6$H$_{13}$	OH	CH$_2$-Ph	1.0
52	i-Pr	OH	CH$_2$-Ph	0.1
53[b]	c-Pr	OH	CH$_2$-Ph	0.1
54	i-Bu	OH	CH$_2$-Ph	1.0
55	c-Hex	OH	CH$_2$-Ph	0.3
56	CH$_2$-c-Pen	OH	CH$_2$-Ph	0.1
57	CH$_2$-c-Hex	OH	CH$_2$-Ph	0.1
58[b]	3-Pyridyl	OH	CH$_2$-Ph	10.0
59	Benzyl	OH	CH$_2$-Ph	0.1
60	4-F-Benzyl	OH	CH$_2$-Ph	0.1
61	CF$_3$	OH	CH$_2$-Ph	0.1
62	CHF$_2$	OH	CH$_2$-Ph	0.1
63	CF$_2$CF$_3$	OH	CH$_2$-Ph	0.3
64	(CF$_2$)$_2$CF$_3$	OH	CH$_2$-Ph	0.3
65	(CH$_2$)$_2$CF$_3$	OH	CH$_2$-Ph	0.1

[a] minimum effective concentration (μM); [b] compounds used in the test set.

**Table 4.3.** Group G3: 2-(alkyl)amino-8-hydroxy-9-benzyladenine derivatives.

entry	R_1	R_2	MEC^a
66	*iso*-Pr	H	1.0
67	*iso*-Bu	H	0.1
68[c]	*s*-Bu	H	0.1
69	cyclohexyl	H	0.1
70	$(CH_2)_2OH$	H	1.0
71	$(CH_2)_3OH$	H	1.0
72	$(CH_2)_4OH$	H	1.0
73	$(CH_2)_3NMe_2$	H	10.0
74	$(CH_2)_2OMe$	H	0.1
75	$(CH_2)_3OMe$	H	0.1
76[b]	Bn	H	0.01
77[b]	Ph	H	1.0
78	$(CH_2)_2Ph$	H	1.0
79	3-MeO-Bn	H	1.0
80	4-MeO-Bn	H	1.0
81	4-Me-Bn	H	0.1
82	4-NMe$_2$-Bn	H	1.0
83	4-F-Bn	H	1.0
84	2-Pyridylmethyl	H	1.0
85	3-Pyridylmethyl	H	1.0
86[b]	4-Pyridilmethyl	H	0.1
87[b]	Bn	Me	10.0
88	-$(CH_2)_2NMe(CH_2)_2$-		1.0
89	-$(CH_2)_2O(CH_2)_2$-		1.0
90	-$(CH_2)_2CHMe(CH_2)_2$-		1.0
91	H	H	10.0
92	CH_3	H	1.0
93	C_2H_5	H	1.0
94	C_3H_7	H	0.1
95	C_4H_9	H	0.1
96	C_5H_{11}	H	0.1

[a] minimum effective concentration (μM); [b] compounds used in the test set; [c] compounds containing a stereogenic centre.

Table 4.4. Group G4: 2-alkylamino-8-hydroxy-9-substituted derivatives.
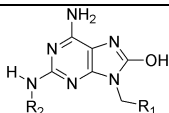
entry	R_1	R_2	MEC[a]
97	2-Cl-Ph	Bu	0.03
98	3-Cl-Ph	Bu	0.03
99	4-Cl-Ph	Bu	0.01
100	2-Me-Ph	Bu	0.03
101	3-Me-Ph	Bu	0.03
102	4-Me-Ph	Bu	0.01
103	4-F-Ph	Bu	0.01
104	4-CF_3-Ph	Bu	0.01
105	4-NO_2-Ph	Bu	0.01
106	4-OMe-Ph	Bu	0.01
107	4-OBnn-Ph	Bu	0.1
108	4-n-Bu-Ph	Bu	0.1
109	4-t-Bu-Ph	Bu	0.03
110	4-Ph-Ph	Bu	0.1
111	4-OH-Ph	Bu	0.01
12b	3,4-Cl_2-Ph	Bu	0.01
113	3,4-F_2-Ph	Bu	0.01
114	2-Cl-4,5-OCH_2O-Ph	Bu	0.1
115	1-naphthyl	Bu	0.01
16b	2-naphthyl	Bu	0.03
17b	5-thiophen(2-Cl)	Bu	0.01
118	CMe_2OH	Bu	0.1
119	i-propyl	Bu	0.1
120	c-hexyl	Bu	0.03
121	2-Py	Bu	0.03
122	3-Py	Bu	0.003
123	4-Py	Bu	0.01
124	2-pyrazine	Bu	0.03
125	4-F-Ph	MeO(CH_2)$_2$	0.03
126	4-F-Ph	$PhCH_2$	0.1
127	4-F-Ph	4-Py-CH_2	0.01
128	2-Me-3-Py	Bu	0.01
129	5-Me-3-Py	Bu	0.003
130	6-Me-3-Py	Bu	0.003
131	2-Cl-3-Py	Bu	0.01
132	4-Cl-3-Py	Bu	0.1
133	6-Cl-3-Py	Bu	0.003
134	6-OMe-3-Py	Bu	0.003
135	6-NMe-3-Py	Bu	0.003
136	2-Cl-6-Me-3-Py	Bu	0.01
137	2,6-Cl-3-Py	Bu	0.03

[a] minimum effective concentration (µM); [b] compounds used in the test set.

Table 4.5. Group G5: 2-alkoxy/2-alkylthio 8-hydroxy-9-benzyladenine derivatives.
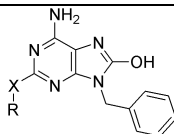
entry	X	R	MEC[a]
138	S	iso-Pr	0.1
139	S	iso-Bu	0.1
140[c]	S	s-Bu	0.05
141	S	cyclohexyl	0.1
142	S	cyclohexylmethyl	0.1
143	S	Ph	1.0
144	S	Bn	0.1
145	S	$(CH_2)_2Me$	0.1
146	S	3-Chlorobenzyl	1.0
147[b]	S	4-Chlorobenzyl	1.0
148	S	3-Methoxybenzyl	0.1
149	S	4-Methoxybenzyl	0.1
150	S	CH_2OMe	0.01
151	S	$(CH_2)_2OMe$	0.1
152	S	$(CH_2)_2OEt$	0.1
153	S	$(CH_2)_3OEt$	0.1
154	S	$(CH_2)_2NMe_2$	10.0
155	S	$(CH_2)_3NMe_2$	0.1
156	S	$(CH_2)_2OH$	0.1
157	S	$(CH_2)_3OH$	0.1
158	S	$(CH_2)_4OH$	0.1
159	O	$(CH_2)_2OMe$	0.01
160	O	$(CH_2)_2OEt$	0.1
161	O	$(CH_2)_3OEt$	1.0
162[b]	O	$(CH_2)_2OH$	1.0
163	O	$(CH_2)_3OH$	1.0
164	S	H	10.0
165	S	CH_3	0.1
166	S	C_2H_5	0.1
167	S	C_3H_7	0.01
168[b]	S	C_4H_9	0.01
169	S	C_5H_{11}	10.0
170	S	$(O)C_4H_9$	0.01
171	O	H	0.1
172	O	CH_3	0.1
173	O	C_2H_5	0.1
174	O	C_3H_7	0.01
175	O	C_4H_9	0.001
176	O	C_5H_{11}	0.01

[a] minimum effective concentration (µM); [b] compounds used in the test set; [c] compounds containing a stereogenic centre.

In 3D-QSAR, a crucial role is played by the relative spatial orientation of structures within the grid.[74] To this purpose, all data set compounds have to be properly aligned. The generation of 3D conformations and alignment of compounds is often a difficult procedure particularly when the ligands are flexible and/or the accurate position/orientation in the binding site is unknown. In our case, experimental bound conformation of ligands (G1-G5 groups) are not yet available since no crystal structure of the putative receptor TLR7 co-crystallized with an IFNI has been reported. To overcome this lack of knowledge, the crystal structure of Imiquimod[163] was used as a starting point for the alignment procedure. First step was to superimpose the most active compound (**175**, MEC = 0.001 μM) together with the most rigid analog (**17**) to the crystal structure of Imiquimod by means of Surflex-Sim[164] which generates molecular alignments and hypothesis of bioactive ligand conformations for 3D approaches. Using Morphological Similarity[165] algorithm based on the shape of molecules, H-bonding and electrostatic properties, Surflex optimizes the alignment of molecules and maximizes their three-dimensional similarity. So aligned Imiquimod, **175** and **17** were therefore used as template for the alignment of the other molecules (Figure 4.2).

The next step was the calculation and application of the molecular interaction fields (MIFs) describing the spatial variation of the interaction energy between a molecular target and a chosen probe. This task was accomplished by the use of the GRID program. Briefly, the protocol involves the computation of MIFs for positions of the probe at points on a rectilinear grid superimposed on the target. Thus, the MIFs provide information on where the favorable and the unfavorable locations for the probe around the target are. To this

---

[163] Zhao, B.; Rong, Y. Z.; Huang, X. H.; Shen, J. S. Experimental and theoretical study on the structure and electronic spectra of imiquimod and its synthetic intermediates. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4942-4946.

[164] Jain, A. N. Ligand-based structural hypotheses for virtual screening. *J. Med. Chem.* **2004**, *47*, 947-961.

[165] Jain, A. N. Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. *J. Comput. Aided. Mol. Des.* **2000**, *14*, 199-213.

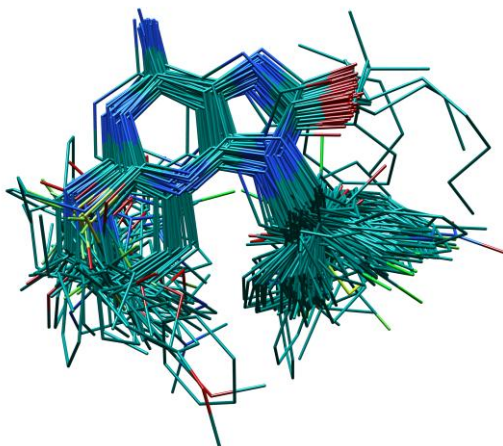aim, OH2, OH, DRY, O, N1 and combined DRY+OH GRID probes
were in turn employed.

Figure 4.2. Stereoview of superimposition of the full data set compounds according
to Surflex-Sim alignment protocol. For the sake of clarity, hydrogen atoms are
omitted.

    The most widespread application of MIFs is likely the derivation
of 3D-QSARs by diverse approaches such as CoMFA or
GRID/GOLPE. Since in 3D-QSAR the biological activity may be
seen as a function of the physicochemical characteristics of the
compounds of interest, the need to convert such numerical data to
useful information has led to the development of methodologies that
rely on statistics and applied mathematics. Among them, an efficient
tool deriving from the Principal Component regression is the Partial
Least Square (PLS) methodology, in which the original variables are
replaced by a small set of linear combinations thereof. Generated
latent variables are then used for multivariate regression in order to
maximize the communality of predictor and response variable
blocks. Additional attractive features of PLS are the ability to
correctly work even when interdescriptor correlations exist and/or to
reduce the risk of chance correlations.

To this point, due to the presence of asymmetric carbon atoms bearing compounds in groups G1, G3 and G5 (Table 4.1, 4.3 and 4.5) our work proceeded considering both *R* and *S* enantiomers. As a consequence, eight training sets (G1-*R*, G1-*S*, G2, G3-*R*, G3-*S*, G4, G5-*R* and G5-*S*) were obtained. The self-consistencies of training sets G1-G5(*R*, *S*) were evaluated developing[166] 48 3D-QSAR sub-models (see models A–AV in Table S7 in the Supporting Information) distinguished by values of correlation coefficient (r^2), cross-validated correlation coefficient (q^2) and cross-validated Standard Deviation Errors of Prediction[167] ($SDEP_{CV}$) falling in the ranges 0.65-0.99, 0.40-0.87, and 0.23-0.56, respectively. Then, to estimate the predictive power of each sub-model, the training sets of the other sub-models were in turn used as external test sets. Thus, whereas a good self-consistency was observed for each sub-model, some of them were affected by a low prediction capability ($SDEP_{ext}$) (see Table S8 in the Supporting Information). The latter results highlight the intrinsic limitation of any statistical modeling or machine learning approach (extrapolation problem) and thus apply also to the above A-AV 3D-QSAR models.

In order to overcome extrapolation problems, two global molecular sets (GMSs) were obtained (GMS-*R* and GMS-*S*) by merging G1-*R*, G2, G3-*R*, G4 and G5-*R* and G1-*S*, G2, G3-S, G4 and G5-*S* series. GMSs were therefore split into a training set of 156 inducers and a test set of 20 inducers (see experimental section). Both training and test sets compounds were selected to cover the range of bioactivities and structures of the GMSs. By the use of aforementioned probes and performing the same procedure, further 12 3D-QSAR models resulted (M1-*R*–M6-*R*, M1-*S*–M6-*S*). In the early stage, the internal predictive power of each individual 3D-QSAR model was checked by 'leave-one-out' (LOO) cross-

---

[166] Ragno, R.; Simeoni, S.; Valente, S.; Massa, S.; Mai, A. 3-D QSAR Studies on Histone Deacetylase Inhibitors. A GOLPE/GRID Approach on Different Series of Compounds. *J. Chem. Inf. Model.* **2006**, *46*, 1420-1430.

[167] Cruciani, G.; Baroni, M.; Clementi, S.; Costantino, G.; Riganelli, D.; Skagerberg, B. Predictive ability of regression models. Part I. Standard deviation of prediction errors (SDEP). *J. Chemom.* **1992**, *6*, 335-346.

validations (CV). The latter indicated that the optimal principal component was the 2nd one for all models except for DRY based ones (see Table 4.6, models M3-*R* and M3-*S*). In fact, cross-validation correlation coefficient q^2 and SDEPs for OH2, OH, N1, O and DRY+OH models showed lower (q^2) and higher (SDEP) values for the 3rd, 4th and 5th principal components. Although, statistical coefficients were not as good as those obtained for the previous 48 sub-models (A-AV in Table S7 in the Supporting Information), the r^2, q^2_{CV-LOO} and SDEP$_{CV-LOO}$ values, included in the ranges of 0.62-0.90, 0.47-0.64 and 0.58-0.71, respectively, can be still considered statistically significant as the q^2_{CV-LOO} values are greater than the threshold values of 0.25, as suggested by Cramer III[168] and in the majority of cases satisfy the more stringent rules depicted by Tropsha.[169] Moreover, different cross-validated PLS analyses were carried out to establish the models robustness (leave-two-out, LTO, leave-some-out, LSO-5, leave-half-out, LHO), confirming the LOO principal components optimal number. It could be argued that the conventional r^2 values for models M1-*R*, M1-*S*, M2-*R*, M2-*S*, M4-*R*, M4-*S*, M5-*R* and M5-*S* are not fully satisfying, ranging between 0.62 (M1-*R*) and 0.73 (M2-*R* and M2-*S*). Nevertheless, by definition of the PLS paradigm, they are highly containing informative and relatively highly robust models in terms of cross validation. As a general trend, the (*R*) configuration training set showed similar or slightly better statistical profiles than models containing the (*S*) enantiomers.

Analyzing values in Table 4.6, it is not surprising that the DRY model shows larger r^2, as the number of latent dimensions is 5 versus 2 for the other models. The DRY model has the largest r^2, the highest dimensionality, but not significantly better q^2 or SDEPs. However, information contained in the DRY based models are of peculiar interest and can be very useful for detecting hydrophobic patches, thus we decided to use it for interpretations.

---

[168] Clark, M.; Cramer, R. D., III. The probability of chance correlation using partial least squares (PLS). *Quant. Struct.-Act. Relat.* **1993**, *12*, 137-145.
[169] Golbraikh, A.; Tropsha, A. Beware of q²! *J. Mol. Graph. Model.* **2002**, *20*, 269-276.

Table 4.6. Statistical results for (M1-*R*–M6-*R*) and (M1-*S*–M6-*S*) 3-D QSAR global models obtained from diverse Golpe PLS analysis.

M	P	r^2	LOO			LTO			LSO-5			LHO			Test set
			q^2	SDEP	PC	q^2	SDEP	PC	q^2	SDEP	PC	q^2	SDEP	PC	$SDEP_{ext}$
M1-R	OH2	0.62	0.47	0.71	2	0.46	0.71	2	0.43	0.73	2	0.39	0.76	2	1.08
M1-S	OH2	0.62	0.48	0.70	2	0.48	0.70	2	0.47	0.70	2	0.44	0.72	2	1.08
M2-R	OH	0.73	0.61	0.61	2	0.61	0.61	2	0.60	0.61	2	0.56	0.64	2	1.05
M2-S	OH	0.73	0.62	0.60	2	0.62	0.60	2	0.61	0.61	2	0.57	0.64	2	1.04
M3-R	DRY	0.89	0.64	0.58	5	0.64	0.58	5	0.61	0.60	5	0.55	0.65	5	0.98
M3-S	DRY	0.90	0.64	0.58	5	0.64	0.58	5	0.61	0.60	5	0.54	0.66	5	1.04
M4-R	N1	0.69	0.58	0.65	2	0.56	0.65	2	0.55	0.65	2	0.52	0.67	2	1.09
M4-S	N1	0.69	0.58	0.64	2	0.56	0.65	2	0.55	0.65	2	0.52	0.67	2	1.06
M5-R	O	0.72	0.60	0.61	2	0.60	0.61	2	0.60	0.62	2	0.57	0.64	2	1.05
M5-S	O	0.69	0.55	0.65	2	0.55	0.65	2	0.55	0.65	2	0.52	0.67	2	1.08
M6-R	DRY	0.70	0.56	0.64	2	0.56	0.64	2	0.55	0.65	2	0.53	0.67	2	1.09
M6-S	+OH	0.69	0.57	0.64	2	0.56	0.64	2	0.56	0.64	2	0.53	0.66	2	1.08

***M**: model name; **P**: GRID probe; **LOO**: Leave One Out Cross-validation; **LTO**: Leave Two Out Cross-validation; **LSO-5**: Leave-Some-Out Cross-validation using 5 groups; **LHO**: Leave Half Out; r^2: conventional square correlation coefficient; q^2: cross-validation correlation coefficient; **SDEP**: cross-validated standard error of prediction; **PC**: optimal number of Principal Components; **SDEP$_{ext}$**: Standard Error of Prediction for the external test set.

Achiral models were also investigated. Compounds **2**, **12**, **68** and **140** were kept out from the training set (152 IFNIs) which was in turn analyzed with OH and DRY GRID probes. Resulting 3-D QSAR models did not show statistically significant differences when compared with models M1-*R*, M1-*S*, M3-*R*, M3-*S*. The omitted compounds were added to the original test set and used for external validation.

It is likely that the tiny differences observed for the (*R*) and (*S*) models stem from the independent generation of conformers/alignments, or maybe slightly different positions of the aligned molecules in the grid. Statistical results for achiral models are detailed in Table S10 in the Supporting Information.

Along with the internal validation (cross validation), the predictive capabilities of models (M1-*R*–M6-*R*) and (M1-*S*–M6-*S*) were estimated by the preselected test set consisting of 20 compounds (see Tables 4.1-4.5). The predictive ability of models containing (*R*) enantiomers showed a slightly better profile than those including (*S*) enantiomers. As represented in Table 4.6, the SDEP$_{ext}$ values were comprised between 0.98 and 1.09. Due to the higher statistical profile and the lower SDEP$_{ext}$, models obtained *via* OH and DRY probes (M2-*R* and M3-*R* in Table 4.6) were selected for a deeper analysis and graphical interpretations.

4.3. Discussion

3D-QSAR models interpretations. One of the most important features of 3D-QSAR is the graphical representation of the model making its interpretation easier. In fact, by using GOLPE software, one may display the final model by several available options. Among them, PLS pseudo coefficients, present field and activity contribution plots are very useful.

The basic plot is the PLS pseudo-coefficients plot which allows the grid points visualization of determined molecules/probe interaction energy level, providing whole training set graphical information. As a result, the contour coefficient maps indicate those

areas in which the model has found a high correlation between the ligand-probe interaction energy and the bioactivity.
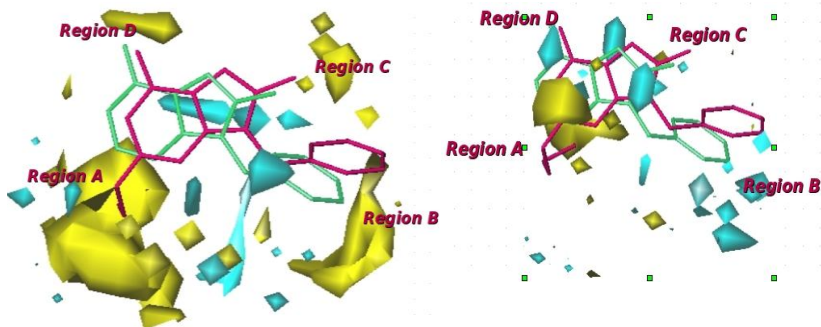
Figure 4.3. GRID/GOLPE PLS Coefficients contour maps for the M2-*R* and M3-*R* 3D-QSAR models (contour levels 0.0049 yellow, -0.0049 cyan; contour levels 0.00452 yellow, -0.00452 cyan, respectively). To aid interpretation only the highest active compound (**175** in red) and one of the lowest active compounds (**36** in green) are shown. For the sake of clarity hydrogen atoms are omitted.

In Figure 4.3 are shown the PLS coefficients plots for models M2-*R* (OH probe) and M3-*R* (DRY probe). The most important areas are located in four regions (herein referred to as A, B, C and D) which encompass the adenine and quinoline scaffolds. In particular, a positive PLS coefficients area (colored in yellow) is located around the C-2 position of the adenine scaffold corresponding to the quinoline 5a carbon atom (region A). Due to the GRID force field characteristics of the OH probe, A reflects areas in which the ligands could make either electrostatic or steric interactions. To properly allocate the nature of region A, an additional 3D-QSAR model was built using the C3 probe (mainly steric information).[71] Statistical results of the latter model (see Table S9 in the Supporting Information) showed to be in good agreement with the previous models. In fact, comparing the C3 probe PLS coefficients with those of M2-*R*, a full overlap was detected (compare Figure S1 in the Supporting Information with Figure 4.3) highlighting as mainly steric or hydrophobic the characteristics of region A. Moreover, for chemical interpretation one must remember that substituents being

able to favorably interact with an OH probe led to negative GRID energies. Hence, cyan contours indicate attractive interactions, which will increase biological affinity, while repulsive, unfavorable interactions will lower activity. In contrast, yellow contours in coefficients maps (see Figure 4.3, OH probe) suggest that repulsive interactions (space filling) increase affinity, while attractive interactions decrease activity. While the PLS coefficients plot provides whole training set graphical information, the activity contribution plots (resulting from the multiplication of PLS coefficients by the actual field values of each object) help focusing on spatial regions that are individually important for the selected molecules. In the bottom of Figure 4.4 (model M2-*R* activity contribution plot for **175**, the most active IFNI), yellow polyhedrons appear in the region surrounding the 2-butyloxy group, suggesting that this substituent may increase IFN-inducing activity. On the other hand, in the activity contribution plot of compound **36**, one of the less active IFNI*s* lacking of any substituents in position C-2, in region A some yellow polyhedrons are converted into cyan colored ones (negative activity contribution). This confirms the steric nature of region A that somehow has to be filled with neutral groups.

Region B is located in proximity of substituents on adenine N-9 or imidazoquinoline N-1 positions. The majority of compounds are characterized by hydrophobic bulky alkyl, aryl or heteroaryl groups at this position. For instance, the most active compound of our training set (**175**) displays a benzyl substituent at this position. As shown in the right side of Figure 4.3, some cyan polyhedrons (negative PLS coefficients) are close to the hydrophobic moieties and, since DRY probe gives rise only to negative field values (Figure S2 in the Supporting Information), the presence of a hydrophobic group in this region positively contributes to the biological activity. Furthermore, to properly explain the influence of the regiochemistry of the substituents at various positions on the benzene ring, we examined PLS-coefficients and activity contribution plots of para-, meta- and ortho-substituted compounds (Figure 4.5). As clearly depicted, para-substituted compound **99** (colored in orange)

displays a yellow polyhedron in proximity of the 4-chloro group even in the PLS coefficients plot or in the activity contributions one (see also Figure 4.6). In contrast, compounds **97** and **98**, respectively, ortho- and meta-chloro substituted, did not orient their groups into this yellow contour (PLS-coefficient plots). Moreover, with reference to the activity contributions plots of both derivatives **97** and **98** (Figure 4.6), it is obvious the deficiency of the yellow polyhedron (positive activity contribute) in the vicinity of the chloro groups, somehow confirming[155] the importance for the biological activity of a correct regiochemistry of the substituents in this area.
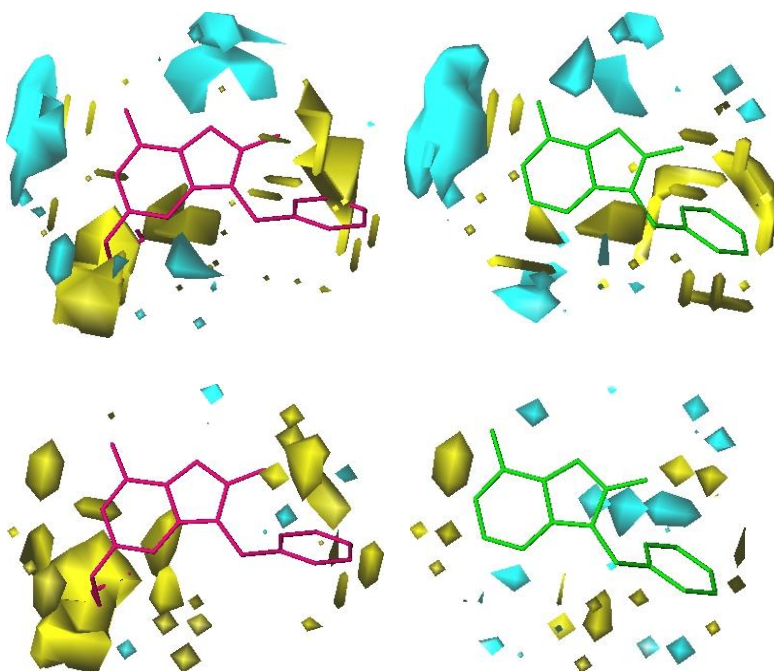
Figure 4.4. On the top, graphical representation of Present Field contour maps (contour levels 3.7787 yellow; -2.3306 cyan) for compounds **175** (in red) and **36** (in green). On the bottom, Activity Contribution plots (contour levels 0.0230 yellow; -0.0230 cyan) for the same compounds.
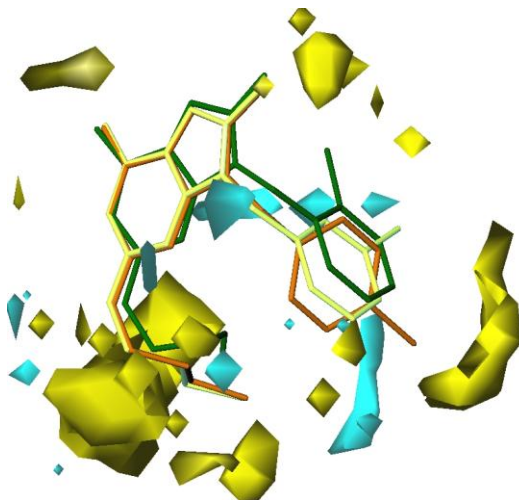
Figure 4.5. GRID/GOLPE PLS Coefficients contour maps for the M2-*R* 3-D QSAR model (contour levels 0.0049 yellow, -0.0049 cyan). Compounds **97**, **98** and **99** are depicted in green, yellow and orange, respectively. For the sake of clarity hydrogen atoms are omitted.
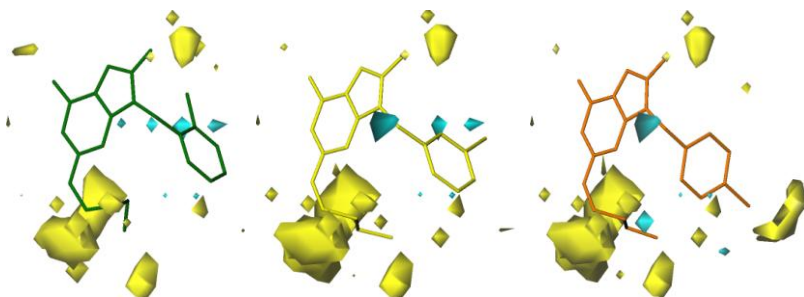
Figure 4.6. Graphical representation of the Activity Contribution plots (contour levels 0.0270 yellow, -0.0270 cyan) for the compounds **97**, **98** and **99** are depicted in green, yellow and orange, respectively. For the sake of clarity hydrogen atoms are omitted.

Regarding regions C and D of Figure 4.3, as previously reported[150,159] an hydroxyl group at adenine C-8 position (region C) and an amino group at adenine C-6 or imidazoquinoline C-4 positions (region D) are essential for IFN-inducing activity. Despite the reported structure-activity relationships (SARs), in the PLS coefficients plots of Figure 4.3 only some small yellow polyhedrons are present in proximity to these positions. However, due to the constant presence of both the C-8 hydroxyl substituent and the C-6/C-4 amino groups the fractional factorial design (FFD) selection led to the exclusion of the majority of grid points around these groups, resulting in a small number of polyhedrons. The fewness of selected lattice points in these areas does not mean that these groups have no influence on the biological activity, but that the majority of compounds included in the training set exert similar influence in that particular area. Nevertheless, since visible in Figure 4.3 only, small polyhedrons suggest the influence of hydrogen bonding in these positions.

4.4. Conclusions

In this work we report the first GRID/GOLPE 3D-QSAR study on a training set of 156 IFN-inducers, including 8-hydroxyadenines and 1*H*-imidazo[4,5-*c*]quinolines derivatives. Twelve 3D-QSAR final models were built and analyzed. All models displayed good statistical coefficient values and using an external test set, model M2-*R* proved to be predictive showing low standard error of predictions. Results from 3D-QSAR model interpretation, let us individuate the most important areas around the molecules and are also in fully agreement with the previously reported SARs. Analysis in terms of principal interaction points between the most actives and the GRID probes let to the definition of a plausible pharmacophoric model (Figure 4.7). To this, around the adenine and quinoline scaffold four regions seem particularly important for the modulation of the IFN inducing activity: (a) a fillable steric pocket (S); (b) a hydrophobic area (Hy); (c) an acceptor hydrogen bonding region close to the adenine C-6/quinoline C4 amino groups (HA) and (d) a

polarized area, possibly with donator hydrogen bonding characteristics close to adenine C-8. These features are fully in agreement with several anti-HCV derivatives able to stimulate interferon release in PBMC (peripheral blood mononuclear cell), recently reported by Czarniecki[170] and Chong in a 2008 patent.[171]
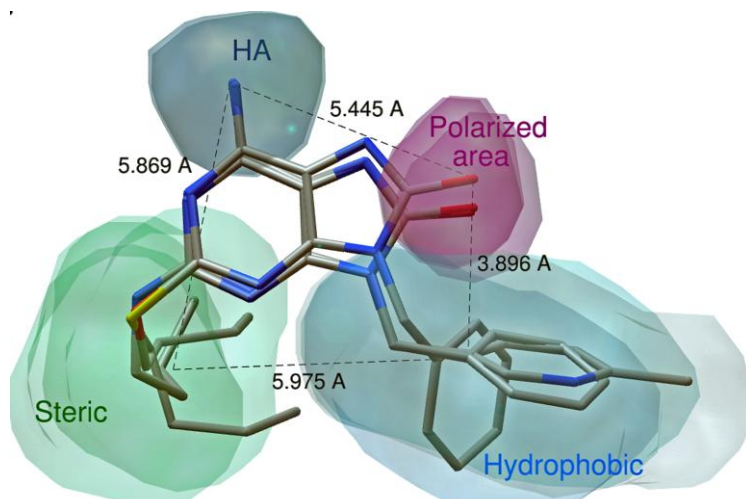
'

Figure 4.7. Graphical representation of the pharmacophoric model obtained from the analysis of the 3D-QSAR models.

4.5. Experimental section

Data sets. For the investigation, 221 IFNIs were selected from literature (see paragraph 4.2). According to the original papers, *in vitro* bioactivities were expressed as minimum effective concentration (MEC, μM) the concentration of compounds required for more than 1 IU/mL induction of interferon. These numbers were then transformed to the decadic log values of the molar MEC which

---

[170] Czarniecki, M. Small Molecule Modulators of Toll-like Receptors. *J. Med. Chem.* **2008**, *51*, 6621-6626.

[171] Chong, L. S.; Desai, M. C.; Gallagher, B.; Graupe, M.; Halcomb, R. L.; Yang, H.; Zhang, J. R. Preparation of purine and thiadeazapurine phosphonate derivatives as modulators of toll-like receptor 7. PCT Int. Appl. WO 2008005555, 2008.

were used as dependent variables in the present 3-D QSAR analyses. All IFNI*s* with not defined MEC values were excluded from the work leading to a global set of 176 compounds (data set, Tables S1-S6 in the Supporting Information).

IFNIs of tables S1-S6 were firstly collected into five distinct groups according to both molecular scaffolds (1*H*-imidazo[4,5-*c*]quinolines and adenines) and substituents at positions 2, 8 and 9 on the adenine ring, providing: 1*H*-imidazo[4,5-*c*]quinolines (group G1), 2-hydrogen/ alkyl/ cycloalkyl/ aryl-8-hydroxy/ mercapto-9-(substituted)arylalkyl adenines (group G2), 2-(alkyl)amino-8-hydroxy-9-benzyladenines (group G3), 2-alkylamino-8-hydroxy-9-(substituted)arylalkyl/ alkyl/ alkylhydroxy (group G4) and 2-alkoxy/ alkylthio-8-hydroxy-9-benzyladenines (group G5). G1-G5 groups chemical structures and biological activities are reported in Table 4.1- 4.5, respectively.

For the final 3D-QSAR models, the training set comprises a series of 156 interferon inducers, whose chemical structures belong to the 8-hydroxyadenines and 1*H*-imidazo[4,5-*c*]quinolines classes. The remaining 20 compounds were used as external test set. Either training or test sets were selected so that they span the whole descriptor (chemical) space occupied by the entire data set and cover the same range of bioactivities. Even if, test set compounds are only 20, they do contain the maximal diversity in terms of molecular structures and bioactivities. All derivatives[159] with activity values reported in *µ*g/mL were transformed in *µ*M to have uniformity of data. Attempts to use concentrations of IFN induced by compounds as dependent variables for our 3-D QSAR models, were unsuccessful (data not shown).

Molecular Modeling and Alignments. All molecules were built starting from ASCII text using the stand-alone version of PRODRG[160,161] in conjunction with GROMACS[162] suite and UCSF CHIMERA[172] as graphical user interface on a 3GHz AMD CPU equipped IBM compatible workstation with the SUSE 9.0 version of

---

[172] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **2004,** *25*, 1605-12.

the Linux operating system. SURFLEX[164,165] software was employed for the alignment procedure. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco.

GRID Calculations. The GRID[93,94] program (version22) was used to describe the aligned molecular structures. Due to the presence of asymmetric carbon atoms in compounds of groups G1, G3 and G5 (Table 4.1, 4.3 and 4.5), both (*R*) and (*S*) enantiomers were considered. As a result, eight initial training sets (G1-*R*, G1-*S*; G2; G3-*R*, G3-*S*; G4 G5-*R* and G5-*S*) were obtained. Since no X-ray data for the TLR7 receptor is still available, a random probe selection was accomplished. Calculations on the 48 obtained sub-models were carried out using probes OH2 (water), OH (hydroxyl group bonded to an aromatic system), DRY (hydrophobic), O (sp^2 carbonyl oxygen), N1 (neutral flat NH) and combined DRY+OH. The energy calculations were implemented using a 1Å spacing between the grid points. Each set of calculated interaction energies contained in the resulting three-dimensional matrix, arranged as one-dimensional vector, was used as input for the program GOLPE. The *xyz* coordinates (in Angstrom) of the grid rectangular box used for the computation are: X_{min}/X_{max}= −6.57/10.95, Y_{min}/Y_{max}= −6.83/13.24, $Z_{min}/Z_{max\ =}$ −10.82/10.19.

The models graphical interpretation was conducted considering OH and DRY probes. To aid graphical interpretation, further 3D-QSAR models were built using the C3 GRID probe.

GOLPE Analyses. PLS models were calculated with GOLPE 4.6.0 running on a PC IBM compatible hardware running the Linux operating system. To measure the goodness of the model the statistical indices r^2, q^2 and SDEP were employed using different cross-validation methods.

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - Y_{fit})^2}$$

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - Y_{cv})^2}$$

$$SDEP = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y_{cv})^2}{N}}$$

where Y_i = experimental value; Y_{fit} = recalculated value; $\overline{Y}$ = mean value; Y_{cv} = predicted value and N = number of experiments.

Variable Preselection. The resulting probe-target interaction energies for each compound were first unfolded to produce one-dimensional vector variables then assembled in the so-called X matrix. The matrix was pretreated using a cutoff of 5 kcal/mol in order to produce a more symmetrical distribution of energy values. To obtain a model with better predictive capability, variable selection was operated by zeroing values with absolute values smaller than 0.01 kcal/mol and removing variables with standard deviation below 0.05. In addition, variables taking only two and three distribution were also removed. Attempts to use other settings in the variable preselection did not lead to any improvement of the models.

Smart Region Definition (SRD). A number of seeds (1000) were selected using a D-optimal design criterion in the weights space. Structural differences between different molecules in the series will be reflected in groups of variables, and therefore groups were generated around each seed in the 3D space. Variables with a critical distance cutoff of 2 Å to the seeds were included in the groups. If two neighboring groups (with a distance smaller than 10 Å) contained the same information the groups were collapsed. The

groups were used in the variable selection procedure replacing the original variables. The effect of the groups on the predictability was evaluated and groups instead of individual variables were removed from the data file. Attempts to use other settings in the SRD did not lead to any improvement of the models.

Variable Selection. The effect of the grouped variables on the predictive ability was evaluated using a fractional factorial design (FFD) procedure. A number of reduced models (twice the number of variables) were built removing some of the variables according to the FFD design. The effect of dummy variables (20%) on the predictive ability was calculated and only variables with a positive effect on the predictability larger than the effect of the average dummy variable were included in the final model. During FFD the predictive ability of the generated matrices was evaluated by cross-validation, using 5 random groups and 100 randomizations. Weights were recalculated after objects exclusion. Variables which were found to be uncertain were retained. The FFD selection was repeated until the r^2 and q^2 value did not increase significantly for all models. In the FFD selection the cross validation was conducted using 5 random groups and a maximum of 5 principal components.

Cross-Validation. All 3D-QSAR models were tested by cross-validation using LOO (leave-one-out), LTO (leave-two-out), LSO-5 (leave-some-out with 5 random groups) and LHO (leave-half-out) methods with 100 randomizations. The optimum number of principal components used in any 3D-QSAR model, was chosen on the basis of the different cross-validated PLS analyses setting to 5 the maximum number of principal components to be extracted. For all 3D-QSAR models in the present study, the optimal number of principal components (PC= 2 for OH2, OH, N1, O and DRY+OH based models and PC= 5 for DRY based model) refers to the ones where $SDEP_{cv}$ assumes the minimum value and the cross-validation correlation coefficient q^2 the maximum value. Attempts to increase the maximum principal components during either CVs or in the above FFD variable selection did not lead to any substantial different model.

4.6. Supporting Information Available

Molecular structures and biological activities of the compounds as reported in the original papers (Table S1-S6), various statistical results (Table S7-S10) and C3 based 3D-QSAR GOLPE contour maps.

**Table S1.** 1H-imidazo[4,5-c]quinolines.
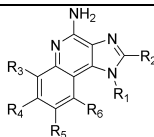
Paper[a]	R_1	R_2	R_3	R_4	R_5	R_6	MEC[b]
Imi	$CH_2CH(CH_3)_2$	H	H	H	H	H	2.08
1	$CH_2CH(OH)CH_2OH$	H	H	H	H	H	9.68
2	CH_3	H	H	H	H	H	2.52
22	CH_2CH_3	H	H	H	H	H	2.36
23	$CH_2CH_2CH_3$	H	H	H	H	H	2.21
24	$CH(CH_3)_2$	H	H	H	H	H	2.21
25	$CH_2CH_2CH_2CH_3$	H	H	H	H	H	2.08
28	$CH_2(CH_2)_3CH_3$	H	H	H	H	H	1.97
29	$CH_2(CH_2)_4CH_3$	H	H	H	H	H	1.86
30	$CH_2(CH_2)_5CH_3$	H	H	H	H	H	1.77
33	CH_2CH_2OH	H	H	H	H	H	2.19
34	$CH_2CH(CH_3)OH$	H	H	H	H	H	1.03
35	$CH_2C(CH_3)_2OH$	H	H	H	H	H	1.95
37	CH_2Ph	H	H	H	H	H	0.36
38	CH_2CH_2Ph	H	H	H	H	H	1.73
40	$CH_2CH_2CH_2CH_2Ph$	H	H	H	H	H	0.32
41	H	H	H	H	H	H	2.71
42	$CH_2CH(CH_3)_2$	H	H	H	H	H	0.20
43	$CH_2CH(CH_3)_2$	CH_2CH_3	H	H	H	H	0.19
44	$CH_2CH(CH_3)_2$	$CH_2CH_2CH_3$	H	H	H	H	0.18
45	$CH_2CH(CH_3)_2$	$CH_2CH_2CH_2CH_3$	H	H	H	H	0.03
46	$CH_2CH(CH_3)_2$	$CH_2(CH_2)_3CH_3$	H	H	H	H	0.16
47	$CH_2CH(CH_3)_2$	$CH_2(CH_2)_4CH_3$	H	H	H	H	1.54
48	$CH_2CH(CH_3)_2$	$CH_2(CH_2)_5CH_3$	H	H	H	H	2.95
50	$CH_2CH(CH_3)_2$	CH_2OPh	H	H	H	H	0.29
52	$CH_2CH(CH_3)_2$	CH_2Ph	H	H	H	H	0.15
53	$CH_2CH(CH_3)_2$	CH_2CH_2Ph	H	H	H	H	2.90
54	$CH_2C(CH_3)_2OH$	CH_3	H	H	H	H	0.18
55	$CH_2C(CH_3)_2OH$	CH_2CH_3	H	H	H	H	0.18
56	$CH_2C(CH_3)_2OH$	$CH_2CH(CH_3)_2$	H	H	H	H	0.03
57	Ph	$CH_2CH_2CH_2CH_3$	H	H	H	H	1.58
66	$CH_2CH(CH_3)_2$	H	OH	H	H	H	3.90
67	$CH_2CH(CH_3)_2$	H	H	CH_3	H	H	1.97
68	$CH_2CH(CH_3)_2$	H	H	OCH_3	H	H	3.70
69	$CH_2CH(CH_3)_2$	H	H	OH	H	H	1.95

[a] compound enumeration as in the original paper; [b] minima effective concentration (μM).

Table S2. Adenine derivatives substituted at position 2, 8 and 9.

Paper no.[a]	R_1	R_2	R_3	MEC[b]
6	H	OH	CH_2Ph	10.0
8	H	SH	CH_2Ph	10.0
16	H	OH	α-naphtylmethyl	10.0
17	H	OH	$CH_2C_6H_4$ (4-F)	10.0
18	H	OH	$CH_2C_6H_4$ (4-Cl)	10.0
19	H	OH	$CH_2C_6H_4$ (4-CH_3)	10.0
20	H	OH	$CH_2C_6H_4$ (2-OCH_3)	10.0
21	H	OH	$CH_2C_6H_4$ (3-OCH_3)	10.0
22	H	OH	$CH_2C_6H_4$ (4-OCH_3)	10.0
24	CH_3	OH	CH_2Ph	1.0
25	C_2H_5	OH	CH_2Ph	1.0
26	C_3H_7	OH	CH_2Ph	1.0
27	C_4H_9	OH	CH_2Ph	0.03
28	Ph	OH	CH_2Ph	10.0
29	SH	OH	CH_2Ph	10.0
30	SCH_3	OH	CH_2Ph	0.10
31	SC_2H_5	OH	CH_2Ph	0.10
32	SC_3H_7	OH	CH_2Ph	0.01
33	SC_4H_9	OH	CH_2Ph	0.01
34	SC_5H_{11}	OH	CH_2Ph	10.0
35	SCH_2Ph	OH	CH_2Ph	0.10
36	$S(O)C_4H_9$	OH	CH_2Ph	0.01
37	NH_2	OH	CH_2Ph	10.0
38	$NHCH_3$	OH	CH_2Ph	1.0
39	NHC_2H_5	OH	CH_2Ph	1.0
40	NHC_3H_7	OH	CH_2Ph	0.10
41	NHC_4H_9	OH	CH_2Ph	0.10
42	NHC_5H_{11}	OH	CH_2Ph	0.10
43	OH	OH	CH_2Ph	0.10
44	OCH_3	OH	CH_2Ph	0.10
45	OC_2H_5	OH	CH_2Ph	0.01
46	OC_3H_7	OH	CH_2Ph	0.01
47	OC_4H_9	OH	CH_2Ph	0.001
48	OC_5H_{11}	OH	CH_2Ph	1.0

[a] compound enumeration as in the original paper; [b] minima effective concentration (μM).

Table S3. 2-alkyl-8-hydroxy-9-benzyl adenine derivatives.
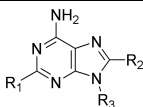
Paper no.[a]	R_1	R_2	R_3	MEC[b]
9e	Pen	OH	CH_2Ph	0.03
9f	Hep	OH	CH_2Ph	1.0
9g	*i*-Pr	OH	CH_2Ph	0.10
9h	*c*-Pr	OH	CH_2Ph	0.10
9i	*i*-Bu	OH	CH_2Ph	1.0
9j	*c*-Hex	OH	CH_2Ph	0.10
9k	CH_2-c-Pen	OH	CH_2Ph	0.10
9l	CH_2-c-Hex	OH	CH_2Ph	10.0
9n	3-Pyridyl	OH	CH_2Ph	0.10
9o	Benzyl	OH	CH_2Ph	0.10
9p	4-F-Benzyl	OH	CH_2Ph	0.10
9q	CF_3	OH	CH_2Ph	0.10
9r	$CHCF_2$	OH	CH_2Ph	0.10
9s	CF_2CF_3	OH	CH_2Ph	0.30
9t	$(CF_2)_2CF_3$	OH	CH_2Ph	0.10
9u	$(CH_2)_2CF_3$	OH	CH_2Ph	0.10

[a] compound enumeration as in the original paper; [b] minima effective concentration (μM).

Table S4. 2-Amino-8-hydroxy-9-Benzyl adenine derivatives.
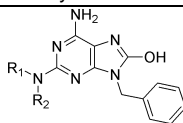
Paper no.[a]	R$_1$	R$_2$	MEC[b]
9f	i-Pr	H	1.0
9g	i-Bu	H	0.10
9h	s-Bu	H	0.10
9i	c-Hex	H	0.10
9j	(CH$_2$)$_2$OH	H	1.0
9k	(CH$_2$)$_3$OH	H	1.0
9l	(CH$_2$)$_4$OH	H	1.0
9n	(CH$_2$)$_3$NMe$_2$	H	10.0
9o	(CH$_2$)$_2$OMe	H	0.10
9p	(CH$_2$)$_3$OMe	H	0.10
9q	Bn	H	0.01
9r	Ph	H	1.0
9s	(CH$_2$)$_2$Ph	H	1.0
9t	3-MeO-Bn	H	1.0
9u	4-MeO-Bn	H	1.0
9v	4-Me-Bn	H	0.10
9w	4-NMe$_2$-Bn	H	1.0
9x	4-F-Bn	H	1.0
9y	2-Pyridylmethyl	H	1.0
9z	3-Pyridylmethyl	H	1.0
9aa	4-Pyridylmethyl	H	0.10
9ab	Bn	Me	10.0
9ac	-(CH$_2$)$_2$NMe(CH$_2$)$_2$-		1.0
9ad	-(CH$_2$)$_2$O(CH$_2$)$_2$-		10.0
9ae	-(CH$_2$)$_2$CHMe(CH$_2$)$_2$-		1.0

[a] compound enumeration as in the original paper; [b] minima effective concentration (μM).

Table S5. 2-alkylthio/2-alkoxy-8-hydroxy-9-benzyl adenine derivatives.
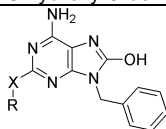
Paper no.[a]	X	R	MEC[b]
11f	S	i-Pr	0.10
11g	S	i-Bu	0.10
11h	S	s-Bu	0.10
11i	S	c-Hex	0.10
11j	S	c-HexMe	0.10
11k	S	Ph	1.0
11m	S	$(CH_2)_2Ph$	0.10
11n	S	3-ClBn	1.0
11o	S	4-ClBn	1.0
11p	S	3-OMeBn	0.10
11q	S	4-OMeBn	0.10
11r	S	CH_2OMe	0.01
11s	S	$(CH_2)_2OMe$	0.10
11t	S	$(CH_2)_2OEt$	0.10
11u	S	$(CH_2)_3OEt$	0.10
11v	S	$(CH_2)_2NMe_2$	10.0
11w	S	$(CH_2)_3NMe_2$	0.10
11x	S	$(CH_2)_2OH$	0.10
11y	S	$(CH_2)_3OH$	0.10
11z	S	$(CH_2)_4OH$	0.10
13f	O	$(CH_2)_2OMe$	0.01
13g	O	$(CH_2)_2OEt$	0.1
13h	O	$(CH_2)_3OEt$	1.0
13i	O	$(CH_2)_2OH$	1.0
13j	O	$(CH_2)_3OH$	1.0

[a] compound enumeration as in the original paper; [b] minima effective concentration (μM).

Table S6. 2-amino-substituted-8-hydroxy-9-substituted adenine derivatives.
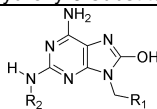
paper no.[a]	R$_1$	R$_2$	MEC[b]
9a	2-Cl-Ph	Bu	0.03
9b	3-Cl-Ph	Bu	0.03
9c	4-Cl-Ph	Bu	0.01
9d	2-Me-Ph	Bu	0.03
9e	3-Me-Ph	Bu	0.03
9f	4-Me-Ph	Bu	0.01
9g	4-F-Ph	Bu	0.01
9h	4-CF$_3$-Ph	Bu	0.01
9i	4-NO$_2$-Ph	Bu	0.01
9j	4-OMe-Ph	Bu	0.01
9k	4-OBnn-Ph	Bu	0.10
9l	4-n-Bu-Ph	Bu	0.10
9m	4-t-Bu-Ph	Bu	0.03
9n	4-Ph-Ph	Bu	0.10
13	4-OH-Ph	Bu	0.01
9o	3,4-Cl$_2$-Ph	Bu	0.01
9p	3,4-F$_2$-Ph	Bu	0.01
9q	2-Cl-4,5-OCH$_2$O-Ph	Bu	0.10
9s	1-naphthyl	Bu	0.01
9t	2-naphthyl	Bu	0.03
9u	5-thiophen(2-Cl)	Bu	0.01
9v	CMe$_2$OH	Bu	0.10
9w	i-propyl	Bu	0.10
9x	c-hexyl	Bu	0.03
9y	2-Py	Bu	0.03
9z	3-Py	Bu	0.003
9aa	4-Py	Bu	0.01
9ab	2-pyrazine	Bu	0.03
20	4-F-Ph	MeO(CH$_2$)$_2$	0.03
21	4-F-Ph	PhCH$_2$	0.10
22	4-F-Ph	4-Py-CH$_2$	0.01
9ac	2-Me-3-Py	Bu	0.01
9ad	5-Me-3-Py	Bu	0.003
9ae	6-Me-3-Py	Bu	0.003
9af	2-Cl-3-Py	Bu	0.01
9ag	4-Cl-3-Py	Bu	0.10
9ah	6-Cl-3-Py	Bu	0.003
9ak	6-OMe-3-Py	Bu	0.003
23	6-NMe-3-Py	Bu	0.003
9ai	2-Cl-6-Me-3-Py	Bu	0.01
9aj	2,6-Cl-3-Py	Bu	0.03

[a] compound enumeration as in the original paper; [b] minima effective concentration (μM).

Table S7. PLS results analyses for G1-*R*-G5-*S* groups using five different single probes and one combined DRY+OH one.

Models	Probes	Groups	q^2	r^2	SDEP	PC
A	OH2	G1-*R*	0.46	0.65	0.44	1
B	OH2	G1-S	0.67	0.93	0.34	3
C	OH2	G2	0.86	0.99	0.35	5
D	OH2	G3-R	0.78	0.98	0.33	5
E	OH2	G3-S	0.77	0.98	0.34	4
F	OH2	G4	0.77	0.99	0.23	5
G	OH2	G5-R	0.54	0.86	0.54	3
H	OH2	G5-S	0.54	0.87	0.54	3
I	OH	G1-R	0.69	0.97	0.34	5
J	OH	G1-S	0.70	0.96	0.33	4
K	OH	G2	0.82	0.99	0.41	5
L	OH	G3-R	0.81	0.99	0.30	5
M	OH	G3-S	0.72	0.95	0.38	3
N	OH	G4	0.76	0.99	0.23	5
O	OH	G5-R	0.57	0.91	0.52	3
P	OH	G5-S	0.54	0.87	0.54	3
Q	DRY	G1-R	0.40	0.76	0.47	2
R	DRY	G1-S	0.44	0.98	0.45	5
S	DRY	G2	0.80	0.98	0.42	4
T	DRY	G3-R	0.72	0.99	0.38	5
U	DRY	G3-S	0.70	0.99	0.41	5
V	DRY	G4	0.62	0.99	0.29	5
W	DRY	G5-R	0.51	0.95	0.56	4
X	DRY	G5-S	0.41	0.95	0.54	4
Y	N1	G1-R	0.68	0.97	0.34	5
Z	N1	G1-S	0.69	0.97	0.34	5
AA	N1	G2	0.87	0.99	0.34	5
AB	N1	G3-R	0.78	0.99	0.34	5
AC	N1	G3-S	0.76	0.95	0.35	3
AD	N1	G4	0.78	0.99	0.23	5
AE	N1	G5-R	0.56	0.87	0.53	3
AF	N1	G5-S	0.54	0.86	0.54	3
AG	O	G1-R	0.64	0.97	0.36	5
AH	O	G1-S	0.50	0.70	0.43	1
AI	O	G2	0.81	0.98	0.41	5
AJ	O	G3-R	0.78	0.99	0.34	5
AK	O	G3-S	0.74	0.95	0.36	3
AL	O	G4	0.73	0.96	0.25	5
AM	O	G5-R	0.62	0.92	0.49	4
AN	O	G5-S	0.66	0.94	0.49	4
AO	DRY+OH	G1-R	0.68	0.98	0.34	5
AP	DRY+OH	G1-S	0.69	0.96	0.34	4
AQ	DRY+OH	G2	0.82	0.99	0.30	5
AR	DRY+OH	G3-R	0.84	0.99	0.29	5
AS	DRY+OH	G3-S	0.79	0.98	0.33	4
AT	DRY+OH	G4	0.77	0.99	0.23	5
AU	DRY+OH	G5-R	0.56	0.92	0.53	4
AV	DRY+OH	G5-S	0.59	0.92	0.50	4

Table S8. PLS results for different probes in order to evaluate
the predictive capability of submodels

H2O					
Training Set R		**SDEPext**			
	G1	G2	G3	G4	G5
G1		0.93	0.82	1.94	1.32
G2	0.81		0.946	1.10	0.93
G3	0.69	1.023		1.32	1.18
G4	1.93	2.056	1.64		1.30
G5	1.58	1.31	1.26	0.72	
Training Set S		**SDEPext**			
	G1	G2	G3	G4	G5
G1		0.81	0.76	1.65	1.13
G2	0.80		0.95	1.10	0.93
G3	0.74	1.08		1.29	1.17
G4	1.93	2.05	1.65		1.30
G5	1.71	1.29	1.31	0.72	

DRY				
Training Set R	**SDEPext**			
	G1	G2	G3	G4	G5
G1		0.86	0.68	1.83	1.25
G2	1.02		0.88	1.94	1.27
G3	0.60	0.98		1.51	1.099
G4	1.66	2.04	1.58		1.16
G5	0.91	1.45	1.23	0.88	
Training Set S	**SDEPext**			
	G1	G2	G3	G4	G5
G1		0.87	0.78	1.87	1.22
G2	1.017		0.92	1.94	1.26
G3	0.60	0.99		1.47	1.05
G4	1.66	2.04	1.58		1.16
G5	1.02	1.46	1.30	0.86	

OH					
Training Set R	**SDEPext**				
	G1	G2	G3	G4	G5
G1		0.89	0.82	1.73	1.15
G2	0.73		0.91	1.17	0.98
G3	0.71	1.037		1.309	1.17
G4	1.88	2.019	1.63		1.30
G5	1.68	1.30	1.32	0.77	
Training Set S	**SDEPext**				
	G1	G2	G3	G4	G5
G1		0.80	0.83	1.62	1.077
G2	0.72		0.916	1.17	0.98
G3	0.75	1.08		1.27	1.14
G4	1.88	2.019	1.63		1.30
G5	1.65	1.307	1.18	0.74	

N1				
Training Set R	SDEP$_{ext}$			
	G1	G2	G3	G4	G5
G1		0.91	0.86	1.80	1.32
G2	0.80		0.96	1.064	0.95
G3	0.71	1.06		1.33	1.106
G4	1.85	2.057	1.63		1.32
G5	1.66	1.30	1.18	0.79	
Training Set S	SDEPext			
	G1	G2	G3	G4	G5
G1		0.89	0.806	1.83	1.16
G2	0.78		0.96	1.064	0.95
G3	0.71	1.13		1.29	1.16
G4	1.84	2.057	1.63		1.32
G5	1.46	1.28	1.08	0.83	

O				
Training Set R	**SDEPext**			
	G1	G2	G3	G4	G5
G1		0.94	0.91	1.92	1.52
G2	0.65		0.88	1.29	1.03
G3	0.81	1.08		1.33	1.16
G4	1.81	2.026	1.59		1.28
G5	1.43	1.37	1.59	0.72	
Training Set S	**SDEPext**			
	G1	G2	G3	G4	G5
G1		0.91	0.90	1.95	1.32
G2	0.646		0.87	1.29	1.03
G3	0.67	1.19		1.44	1.21
G4	1.808	2.026	1.7		1.28
G5	1.56	1.40	1.62	0.68	

DRY+OH				
Training Set R	**SDEPext**			
	G1	G2	G3	G4	G5
G1		0.87	0.85	1.79	1.27
G2	0.74		0.91	1.16	0.99
G3	0.75	1.038		1.32	1.17
G4	1.87	2.018	1.61		1.3
G5	1.77	1.31	1.35	0.72	
Training Set S	**SDEPext**			
	G1	G2	G3	G4	G5
G1		0.79	0.85	1.67	1.14
G2	0.73		0.91	1.16	0.99
G3	0.74	1.067		1.29	1.13
G4	1.87	2.018	1.62		1.30
G5	1.60	1.28	1.256	0.74	

Table S9. Statistical results of the 3D-QSAR models obtained using C3 GRID probe.

Absolute configuration	r^2	q^2	$SDEP_{cv}$	$SDEP_{ext}$	PC
R	0.67	0.53	0.67	1.2	2

INTERFERON INDUCERS

Table S10. Statistical results of the 3-D QSAR analyses for the achiral models with OH and DRY GRID probes. External validation was carried out with a test set of 24 compounds.

P	r^2	LOO			LTO			LSO-5			LHO			$SDEP_{ext}$
		q^2	SDEP	PC	q^2	SDEP	PC	q^2	SDEP	PC	q^2	SDEP	PC	
OH	0.72	0.60	0.61	2	0.60	0.62	2	0.57	0.43	2	0.54	0.66	2	1.1
DRY	0.90	0.62	0.59	5	0.62	0.60	5	0.60	0.62	5	0.52	0.67	5	1.02

* **P**: GRID probe; **LOO**: Leave One Out Cross-validation; **LTO**: Leave Two Out Cross-validation; **LHO**: Leave Half Out; **LSO-5**: Leave-Some-Out Cross-validation using 5 groups; q^2: conventional square correlation coefficient; r^2: conventional square correlation coefficient; q^2: cross-validation correlation coefficient; **SDEP**: cross-validated standard error of prediction; **PC**: optimal number of Principal Components; **$SDEP_{ext}$**: Standard Error of Prediction for the external test set.

Figure S1.GRID/GOLPE PLS Coefficient contour maps for the *C3 probe* 3-D QSAR model (contour levels 0.0049 yellow, -0.0049 cyan. To aid interpretation only the highest active compound (**175** in red) and one of the lowest active compounds (**36** in green) are shown. For the sake of clarity hydrogen atoms are not shown.
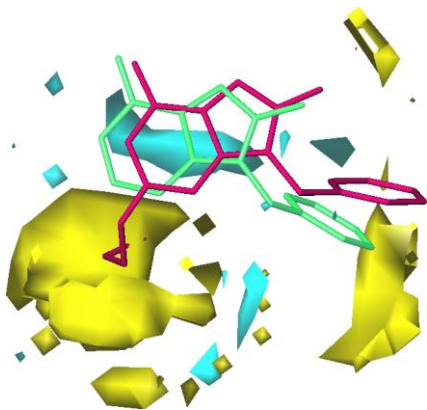
Figure S2.Graphical representation of Present Field contour maps (contour levels -2.5506 cyan).
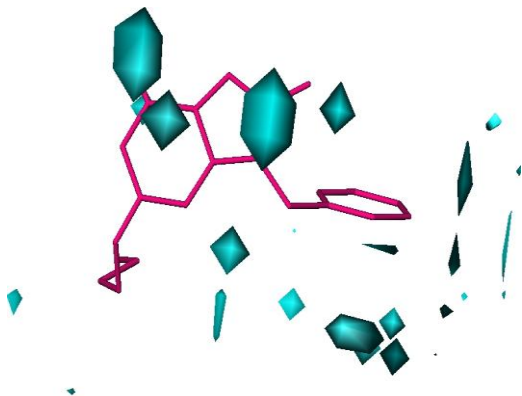
5. NS5B RNA-DEPENDENT RNA POLYMERASE

The remarkable clinical success of human immunodeficiency virus reverse transcriptase and protease inhibitors, as well as the availability of several crystal structures, has motivated HCV drug discovery efforts to focus mainly on the development of protease and polymerase inhibitors. HCV NS5B is an RNA-dependent RNA polymerase that is responsible for the replication of the viral genome. *In vitro*, NS5B can initiate RNA synthesis either by primer-independent initiation, using a single nucleotide complementary to the base at the 3' end of the viral genome (also referred as *de novo* initiation), or by primer-dependent initiation using either DNA or RNA as primers.[173] As is the case for other members of the *Flaviviridae* family of viruses[174], *de novo* initiation of RNA synthesis is thought to be the mechanism used by the polymerase for initiation of RNA synthesis in HCV infected cells.[175]

The HCV RdRp is a "tail-membrane-anchored protein", *i.e.*, membrane association is mediated by a C-terminal 21 amino acid hydrophobic α-helix. The presence of the C-terminal membrane-anchor is necessary for the assembly of a functional replication complex on the ER of infected cells. However, the same region is fully dispensable for *in vitro* polymerase activity. The crystal studies of HCV polymerase from genotype 1 lacking the 21 or 55 C-terminal residues (ΔC21 and ΔC55), were solved initially in the absence of bound substrate or template (apoenzyme).[176,177]

The HCV polymerase structure shares the same general right-handed configuration consisting of finger, thumb, and palm domains

[173] Zhong, W.; Uss, A. S.; Ferrari, E.; Lau, J. Y.; Hong, Z. De novo initiation of RNA synthesis by hepatitis C virus non-structural protein 5B polymerase. *J. Virol.* **2000**, *74*, 2017-2022.

[174] Ortin, J.; Parra, F. Structure and function of RNA replication. *Annu. Rev. Microb.* **2006**, *60*, 305-306.

[175] Butcher, S. J.; Grimes, J. M.; Makeyev, E. V.; Bamefod, D. H.; Stuart, D. I. A mechanism for initiating RNA-dependant RNA polymerazation. *Nature*, **2001**, *410*, 235-240.

[176] Lesburg, C. A.; Cable, M. B.; Ferrari, E.; Hong, Z.; Mannarino, A. F.; Weber, P. C. Crystal structure of the RNA-dependent polymerase from hepatitis C virus reveals a fully encircled active site. *Nat. Struct. Biol.* **1999**, *6*, 937-943.

[177] Bressanelli, S.; Tomei, L.; Roussel, A.; Incitti, I.; Vitale, R. L.; Mathieu, M.; De Francesco, R.; Rey, F. A. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Proc. Natl. Acad. Sci.* U. S. A. **1999**, *96*, 13034-13039.

(Figure 5.1), and the same catalytic machinery observed in DNA-polymerases and HIV-RT. The active site of the enzyme is located in the palm domain that is structurally conserved in all polymerases. Morevover, it is defined by the Asp-$(Xaa)_4$-Asp and the signature GDD motives typical of most polymerases.
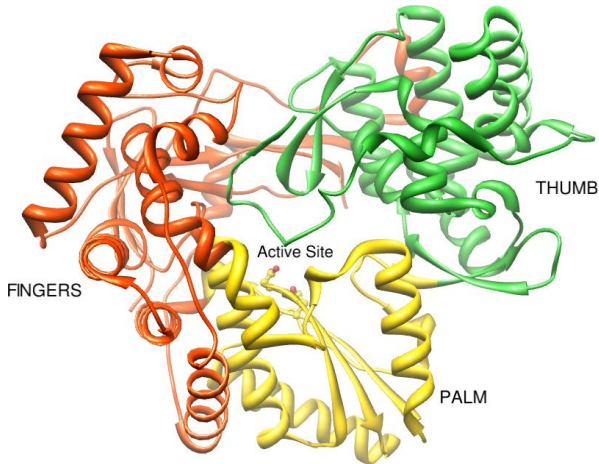
Figure 5.1. Ribbon representation of the HCV NS5B polymerase with domains colored according to thumb (green), palm (yellow), and fingers (orange).

The carboxylate side chains of the first aspartate of each motif coordinate two Mg^{2+} ions involved in the catalytic phosphor-transfer reaction. Compared to DNA polymerases, NS5B has two elongated, flexible loops (Λ1 and Λ2) that extend from the finger domain to the thumb domain resulting in a relatively closed spherical appearance for the enzyme (Figure 5.2).[176,177] The RNA-template and nucleoside triphosphate (NTPs) access the active site, located on the palm domain, through two positively charged tunnels and are then positioned for the reaction through interactions with the thumb and finger domains. Further, a β-hairpin in the thumb subdomain (also referred as β-flap) protrudes toward the enzyme active cavity and has been proposed to act as a 'gate' which monitors the correct binding of the 3'-terminus of the RNA, assuring correct initiation and

replication from the terminus of the HCV genome.[178,179] Crystal structures of the NS5B polymerase bound to a short RNA-template[180] or to NTPs[181] have also been solved providing insights into template binding and mechanism of *de novo* initiation reaction.

In these structures, RNA binding occurred without rearrangements in the enzyme domains. Interestingly, the structure of the NS5B-GTP complex has also showed that, in addition to the active site, GTP can also bind to an allosteric site on the surface of the thumb domain, 30 Å away from the polymerase catalytic center, and in close proximity to a lipophilic region involved in the interaction to the tip of the fingertip Λ1 loop. This external GTP binding site is thought to be an allosteric regulator of finger-thumb interactions and was proposed to exert a regulatory activity by modulating the interactions between fingers and thumb during RNA synthesis.[181,182] However, a role for the external GTP in the NS5B enzymatic activity has not been firmly demonstrated. The essential role of NS5B for viral replication and the success obtained in the clinic with inhibitors of HIV reverse transcriptase (RT) and hepatitis B virus (HBV) DNA polymerase[183,184] have prompted many laboratories to search for inhibitors of the HCV polymerase that could be used in the therapeutic intervention against the virus.

[178] Hong, Z.; Cameron, C. E.; Walker, M. P.; Castro, C.; Yao, N.; Lau, J. Y.; Zhong, W. A novel mechanism to ensure terminal initiation by hepatitis C virus NS5B polymerase. *Virology* **2001**, *285*, 6-11.

[179] Zhong, W.; Ferrari, C. A.; Lesburg, C. A.; Maag, D.; Ghosh, S. K.; Cameron, C. E.; Lau, J. Y.; Hong, Z. Template/primer requirements and single nucleotide incorporation by hepatitis C virus nonstructural protein 5B polymerase. *J. Virol.* **2000**, *74*, 9134-9143.

[180] O'Farrell, D.; Trowbridge, R.; Rowlands, D.; Jager, J. Substrate complexes of hepatitis C virus RNA polymerase (HC-J4): structural evidence for nucleotide import and de-novo initiation. *J. Mol. Biol.* **2003**, *326*, 1025-1035.

[181] Bressanelli, S.; Tomei, L.; Rey, F. A.; De Franesco, R. Structural analysis of the hepatitis C virus RNA polymerase in complex with ribonucleotides. *J. Virol.* **2002**, *76*, 3482-3492.

[182] Lohmann, V.; Overton, H.; Bartenschlager, R. Selective stimulation of hepatitis C virus and pestivirus NS5B RNA polymerase activity by GTP. *J. Biol. Chem.* **1999**, *274*, 10807-10815.

[183] El Safadi, Y.; Vivet-Boudou, V.; Marquet, R. HIV-1 reverse transcriptase inhibitors. *Appl. Microbiol. Biotechnol.* **2007**, *75*, 723-737.

[184] Rivkin, A. A review of entecavir in the treatment of chronic hepatitis B infection. *Curr. Med. Res. Opin.* **2005**, *21*, 1845-1856.
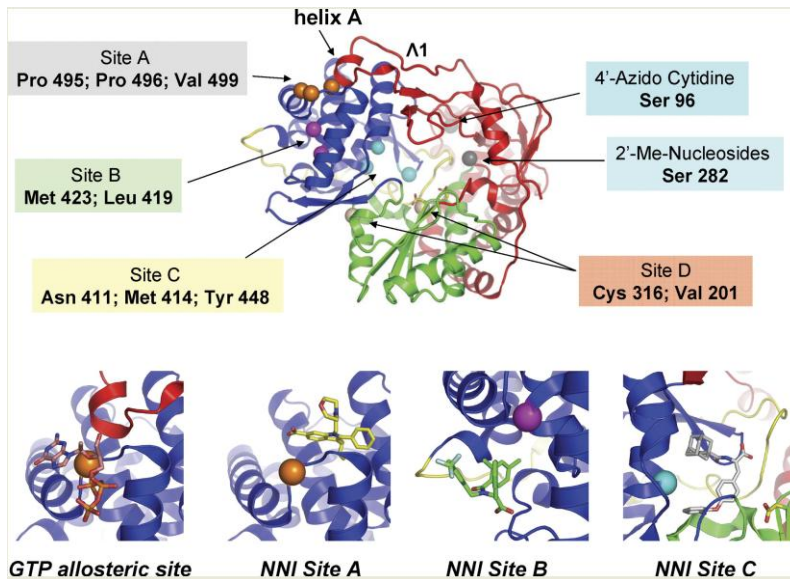
Figure 5.2. Three-dimensional structure of the HCV NS5B RNA-dependent RNA polymerase and non-nucleoside inhibitor binding sites. The finger, palm, and thumb domains are coloured in red, green and dark blue respectively. The panels in the lower part of the figure show a close up view of the allosteric GTP binding site;[185] NNI site A with a bound indole-acetamide inhibitor (in yellow); NNI site B with a bound diphenylalanine inhibitor (in green); NNI site C with a bound phenyl-acylic acid inhibitor (in orange).

5.1. Inhibitors of the NS5B polymerase

The majority of approved antiviral drugs exploit the inhibition of viral polymerase as the primary mechanism of action.[186] Therefore, it is not surprising that a major effort by several pharmaceutical companies has focused on identifying novel inhibitors of the HCV NS5B polymerase. Existing inhibitors of viral polymerases can be tentatively classified into three categories according to their

---

[185] De Francesco, R.; Carfi, A. Advances in the development of new therapeutic agents targeting the NS3-4A serine protease or the NS5B RNA-dependent RNA polymerase of the hepatitis C virus. *Adv. Drug Del. Rev.* **2007**, *59*, 1242-1262.

[186] De Clercq, E. Antiviral drugs: current state of the art. *J. Clin. Virol.* **2001**, *22*, 73-89.

chemical structure and their mechanism of action: (i) nucleoside (substrate) analogues, (ii) non-nucleoside inhibitors (NNI), and (iii) pyrophosphate mimics.

Nucleoside analogues are substrate analogues that need to be phosphorylated to their corresponding nucleoside triphosphate (nucleotide) in the cytoplasm of infected cells in order to become active against the viral polymerases. The nucleotide may be incorporated by the polymerase during processive nucleic acid synthesis, leading to early termination of the elongation reaction and thus inhibition of the virus life cycle. Nucleoside inhibitors of the viral polymerase are used therapeutically for HIV, hepatitis B and herpes viruses. A few different classes of nucleoside analogues have been reported to inhibit HCV replication and, as of today, two investigational agents, valopicitabine (Idenix/Novartis) and R1626 (Roche) have shown anti-HCV activity as single agents in clinical trials.

The NNI class comprises structurally heterogeneous compounds, which usually bind to a site on the enzyme surface away from the enzyme active site, such as an allosteric site. Even if a very large number of HCV polymerase NNI classes have been reported, only one of them, HCV-796 (Viropharma/Wyeth) has displayed significant antiviral activity in HCV-infected patients. In the following sections, different classes of HCV NS5B inhibitors, their mechanism of action, resistance and clinical relevance will be described.

5.1.1.   Nucleoside Analogues

Nucleoside analogues can be considered as prodrugs. These inhibitors are normally synthesized in the non-phosphorylated form in order to facilitate their access into the cell, and are then converted in the active triphosphate (NTP) form by cellular kinases in the cytoplasm of infected cells.[187] If the newly generated NTP analogue is accepted as a substrate by the HCV NS5B polymerase, it can be incorporated in the nascent RNA chain. After insertion of the

---

[187] Carroll, S. S.; Olsen, D. B.; Nucleoside analog inhibitors of hepatitis C virus replication. Infect. *Disord. Drug. Targets*. **2006**, *6*, 17-29.

modified nucleotide, the particular structure present on the sugar moiety will prevent further addition of subsequent bases resulting in the premature arrest of the elongation of the growing polymer chain. Nucleoside analogues that act according to this mechanism of action are therefore referred to as "chain terminators".[188] Bioavailability, efficiency of conversion to the active form by cellular kinases, competition with the natural NTP pool, as well as metabolic deamination and cleavage of the glycosidic bond are all factors to be specifically taken in account for the development of this class of inhibitors.[189] For this reason, development of NIs for HCV has required complex structure-activity relationship (SAR) studies aimed at simultaneous and convergent optimization of the potency of the corresponding NTPs in polymerase enzymatic assays and in cell based replicon or infectivity assays, on one hand, and the pharmacokinetic and metabolic properties, on the other hand.[43] Deoxy-ribonucleoside analogues have been used in the past to target HIV-RT and hepatitis B virus and herpes viruses DNA polymerases. However, the HCV NS5B enzyme is an RNA-dependent RNA-polymerase so is surprising that all the HCV NS5B NIs are considered ribonucleoside analogues where chain termination function has been achieved by relatively small modifications at 2' and 4' position of the ribose. The advantage of such molecules is that they are not expected to be recognized by the endogenous DNA polymerases, avoiding therefore the potential for cytotoxic and genotoxic side effects.

5.1.1.1.   2'-modified nucleoside analogues

2'-Methyl nucleosides have been identified by two independent groups as specific inhibitor of HCV NS5B-dependent RNA

---

[188] Migliaccio, G.; Tomassini, J. E.; Carroll, S. S.; Tomei, L.; Altamura, S.; Bhat, B.; Bartholomew, L. et al. Characterization of resistance to non-obligate chain-terminating ribonucleoside analogs that inhibit hepatitis C virus replication *in vitro*. *J. Biol. Chem.* **2003**, *278*, 49164-49170.

[189] Pierra, C.; Amador, A.; Benzaria, S.; Cretton-Scott, E.; D'Amours, M.; Mao, J.; Mathieu, S.; Moussa, A.; Bridges, E. G.; Standring, J. P.; Sommadossi, R.; Storer, R.; Gosselin, G. Synthesis, and pharmacokinetics of valopicitabine (NM283), an efficient prodrug of the potent anti-HCV agent 2'-C-methylcytydine. *J. Med. Chem.* **2006**, *49*, 6614-6620.

synthesis.[187] Researchers at Idenix reported the nucleoside analogue 2'-C-methyl-cytidine as an inhibitor of HCV replication.[189] This compound was initially identified as an inhibitor of the HCV-related bovine viral diarrhea virus (BVDV) and later shown to have modest activity in the HCV replicon system. Valopicitabine, (Chart 5.1) formerly known as NM283, is an orally bioavailable 3'-L-valine ester of 2'-C-methyl-cytidine,[189,190] that has progressed to advanced clinical trials, in collaboration with Novartis, for the treatment of chronic HCV infection.

In early phase I/IIa clinical trials in patients that included treatment naïve as well as non-responders to previous treatments, valopicitabine monotherapy was shown to reduce viral load of HCV-infected patients in a dose-dependent manner during 15 day treatment.[191] Following these preliminary results and in order to enhance its antiviral potential, valopicitabine was dosed in combination with PEG-IFN-α-2a in patients chronically infected with genotype 1 HCV who had not responded to the previous treatments with the standard PEG-IFN/HCV therapy as well as in treatment-naïve patients. In the course of phase II combination clinical trials, the valopicitabine dose had to be decreased because of severe gastrointestinal side effects observed in some patients at higher doses.

The next steps in the valopicitabine development program included a ribavirin/valopicitabine drug-drug clinical interaction study preliminary to the investigation of triple PEG-IFN-α-2a/ribavirin/valopicitabine in treatment-naïve patients. This study was necessary in light of the finding that, in cell culture, addition of ribavirin antagonizes the anti-HCV activity of the valopicitabine parent compound, 2'-methylcytidine. In this respect, it has been recently disclosed that no negative drug interaction was observed between valopicitabine and ribavirin after 12 weeks of dosing in a

---

[190] Pierra, C.; Benzaria, S.; Amador, A.; Moussa, A.; Mathieu, S.; Storer, R.; Gosselin, G. NM283, an efficient prodrug of the potent anti-HCV agent 2'-C-methylcytidine. *Nucleosides. Nucleotides. Nucleic. Acids.* **2005**, *24*, 767-770.

[191] Afdhal, N.; Godofsky, E.; Dienstag, J.; Rustgi, V.; Schick, L.; McEniry, D.; Zhou, X. J.; Chao, G.; Fang, C.; Fielman, B.; Myers, M. *Hepatology* **2004**, Abs 726 LB703.

trial comparing the standard of care to valopicitabine/ribavirin/PEG-IFN triple therapy.

2'-C-methylcytidine (NM-107)
Enzyme IC_{50} (TP) 1.2 μM
Replicon EC_{50} 25 μM

Valopicitabine

7-Deaza-2'-C-methyladenosine
Enzyme IC_{50} (TP) 0.07 μM
Replicon EC_{50} 0.3μM

2'-deoxy-2'-fluoro-2'-C-methylcytidine (PSI-6130)
Enzyme K_i (TP)   4.3 μM
Replicon EC_{90} 5.4 μM

4'-Azido-cytidine (R1479)
Enzyme IC_{50} (TP) 0.04 μM
Replicon EC_{50} 1.3 μM

R1626
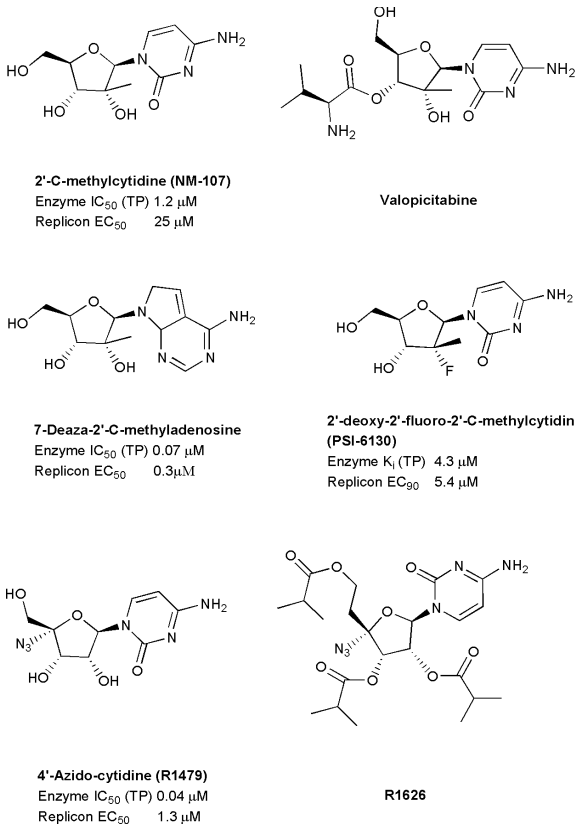
Chart 5.1. Chemical structures of some selected nucleoside inhibitors of the NS5B RdRp.

Other 2'-methyl nucleosides that selectively inhibit HCV replication are 2'-O-methyl-cytidine, 2'-C-methyl-adenosine, 2'-C-methyl-guanosine, and 7-deaza-2'-C-methyl-adenosine (Chart 5.1). These compounds are clinically less advanced, but some of them have been extensively characterized *in vitro* as well as in preclinical models of HCV infection.[192,193] In particular, the high inhibitory potency and reduced cellular toxicity of 7-deaza-2'-C-methyl-adenosine, combined with a very promising pharmacokinetic profile in preclinical animal species, makes this compound an attractive candidate for clinical development.

Lastly, R7128 (Roche/Pharmaset) is a 2'-modefied nucleoside analogue that has recently entered early clinical trials. R7128 is an oral prodrug of 2'-deoxy-2'-fluoro-2'-C-methyl-cytidine (PSI-6130; Chart 5.1), a pyrimidine nucleoside analogue that was shown to act as a non-obligate chain terminator of the NS5B polymerase activity and to efficiently inhibit the replication of HCV RNA in hepatoma cell lines. Data on the in vivo antiviral activity of R1728 are not yet available.

5.1.1.2. 4'-modified nucleoside analogues

A 4-modified nucleoside, 4'-azido-cytidine (R1479, Roche; Chart 5.1), has been reported as a potent competitive inhibitor of NS5B-dependent RNA synthesis and hepatitis C virus replication in cell culture. This compound was shown to inhibit RNA synthesis in a competitive manner and to be a moderately potent inhibitor of the replicon system. A triacylated oral prodrug of 4'-azidocytidine (R1626; Chart 5.1) was dosed in a multiple ascending dose study to patients with chronic genotype 1 infection and was well tolerated, with reversible mild to moderate haematological changes observed

---

[192] Eldrup, A. B.; Allerson, C. R.; Bennett, C. F.; Bera, S.; Bhat, B.; Bhat, N.; Bosserman, J.; Brooks, J. et al. Structure-activity relationship of purine ribonucleosides for inhibition of HCV RdRp. *J. Med. Chem.* **2004**, 47, 2283-2295.

[193] Eldrup, A. B.; Prhavc, M.; Brooks, J.; Bhat, B.; Prakash, T. P.; Song, Q.; Bera, S.; Bhat, N.; Dande, P. et al. Structure-activity relationship of heterobase-modified 2'-C-methyl ribonucleosides as inhibitors of HCV RNA replication. *J. Med. Chem.* **2004**, *47*, 5284-5297.

with increased doses. R1626 will enter a phase II trial in combination with PEG-IFN with or without ribavirin.[194]

5.1.2. Non-nucleoside inhibitors of NS5B

Screening of inhibitor libraries has resulted in the discovery of several classes of allosteric non-nucleoside inhibitors (NNIs).[188,195] The hallmark of all the allosteric HCV NNIs described so far is that – in contrast to active site nucleoside inhibitors – they are non-competitive with NTPs and inhibit the polymerase at a stage preceding the elongation reaction. Despite this apparently common features in the mechanism of inhibition, several different allosteric binding sites for non-nucleoside inhibitors have been discovered on the NS5B RdRp surface. A combination of crystallographic, biochemical and mutagenesis studies have been exploited to identify the binding sites of several of these molecules. For some compounds, structural studies have also allowed to elucidate their mechanism of inhibition. Moreover, the use of these inhibitors in the replicon system allowed selection of clones with resistance mutations mapping in the NS5B polymerase. Of note, the identity and location of the mutations in the polymerase are largely dependent on the structural class of the inhibitor and on their binding site and, more importantly, the mutations obtained against inhibitors binding to one site generally do not confer resistance to compounds binding to a different site.[196] This suggests that NS5B inhibitors with different binding sites could potentially synergize if used in combination and initial biochemical and replicon studies seem to confirm this hypothesis.[197] The presence of several different binding

---

[194] Roberts, S.; Cooksley, C.; Dore, G. et al. Results of phase IB, multiple dose study of R1626, a novel nucleoside analogue targeting HCV polymerase in chronic HCV genotype 1 patients. *Hepatology* **2002**, *44*, 692A.

[195] Koch, U.; Narjes, F.; Allosteric inhibition of the hepatitis C virus NS5B RNA dependent RNA polymerase. *Infect. Disord. Drug. Targets.* **2006**, *6*, 31-41.

[196] Tomei, L.; Altamura, S.; Paonessa, G.; De Francesco, R.; Migliaccio, G. HCV antiviral resistance: the impact of *in vitro* studies on the development of antiviral agents targeting the viral NS5B polymerase. *Antivir. Chem. Chemother.* **2005**, *16*, 225-245.

[197] Wyles, D. L.; Kaihara, K. A.; Vaida, F.; Schooley, R. T. Synergy of small molecular inhibitors of HCV replication directed at multiple viral targets. *J. Virol.* **2007**, *81*, 3005-3008.

sites for small molecules inhibitors in the HCV polymerase is unprecedented and is surprising especially if NS5B is compared to HIV1 RT where only one binding site for NNIs and one for nucleoside analogues have been described.

To date, at least four distinct allosteric binding sites have been described (Figure 5.2). Two of them, sites A and B, are located on the thumb domain whereas the other two, sites C and D, are close to the active site cavity and contact amino acids from both thumb and palm domains (Figure 5.2).

5.1.2.1.   NNIs binding to site A

NNI site A is located in the thumb domain at 30 Å from the active site (Figure 5.2). Several structurally related inhibitors have been shown to bind to this site.[198] Compounds with a benzimidazole scaffold (Chart 5.2) were initially disclosed by Japan Tobacco and Boehringer Ingelheim. Some of the initial molecules were shown to have activity in the cell based assays, with low cytotoxicity, and to be selective versus DNA polymerases.[199,200] Two compounds of this structural family, JTK-109 and JTK-003 were the first HCV NNIs to enter clinical trials, but have now been abandoned. Later compounds with the structurally related indole-6-carboxylic acid (Chart 5.2) quinoline, thienoimidazole-5-carboxylic acid and thienopyrrole scaffolds have been described.[201]

---

[198] Beaulieu, P. L. Finger loop inhibitors of the HCV NS5B polymerase: discovery and prospects for new HCV therapy. *Curr. Opin. Drug Discov.* **2006**, *9*, 618-626.

[199] Hirashima, S.; Suzuki, T.; Ishida, T.; Noji, S. et al Benzimidazole derivatives bearing substituted biphenyls as hepatitis C virus NS5B RNA-dependent RNA polymerase inhibitors: structure-activity relationship studies and identification of a potent and highly selective inhibitor JTK-109. *J. Med. Chem.* **2006**, *49*, 4721-4736.

[200] Ishida, T.; Suzuki, T. et al. Benzimidazole inhibitors of hepatitis C virus NS5B polymerase: identification of 2-[(4-diarylmethoxy)phenyl]-benzimidazole. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1859-1863.

[201] Koch, U.; Narjes, F. Recent progress in the development of inhibitors of the hepatitis C virus RNA-dependent RNA polymerase. *Curr. Top. Med. Chem.* **2007**, *7*, 1302-1329.
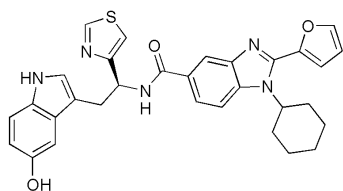
Benzimidazole derivative
Enzyme IC$_{50}$ 0.3 μM
Replicon IC$_{50}$ 1.7 μM
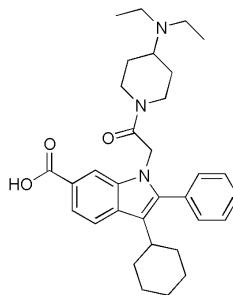
Indole derivative
Enzyme IC$_{50}$ 0.009 μM
Replicon IC$_{50}$ 0.13 μM

Chart 5.2. Chemical structures of site-A non-nucleoside inhibitors of the NS5B RdRp.

The indole and thienopyrrole resulted almost equally potent in the cell based assay and 2 orders of magnitude more potent that the corresponding benzimidazole presumably due to the higher hydrophobicity of their scaffolds.[202,203,204] Site A-NNIs have provided encouraging pre-clnical results as anti-HCV agents. In particular indole-acetamide based inhibitors were shown to be devoid of off-target activities and presented encouraging pharmacokinetic profile in preclinical animal studies. Some of the compounds belonging to the classes just described have very potent inhibitory activity on the genotype 1b and 3a HCV polymerases. On the contrary, all site A inhibitors described to date afford only marginal or no inhibition of the genotype 2 enzyme, presumably due to structural differences in the binding site that reduce the binding activity.

---

[202] Harper, S.; Pacini, B.; Avolio, S.; Di Filippo, M.; Migliaccio, G. et al. Development and preliminary optimization of indole-N-acetamide inhibitors of hepatitis C virus NS5B polymerase. *J. Med. Chem.* **2005**, *48*, 1314-1317.

[203] Beaulieu, P. L.; Gillard, J.; Bykowski, D.; Brochu, C.; Dansereau, N. et al. Improved replicon cellular activity of non-nucleoside allosteric inhibitors of HCV NS5B polymerase: from benzimidazole to indole scaffolds. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4987-4993.

[204] Ontoria, J. M.; Martin Hernanado, J. I.; Malacona, S.; Atteni, B. et al. Identification of thieno[3,2-b] pyrroles as allosteric inhibitors of epatiti C virus NS5B polymerase. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4026-4030.

A high-resolution crystal structure of the NS5B polymerase bound to an indole-acetamide NNI has been reported.[205] The indole and benzimidazole binding site revealed by the structure is consistent with the location of the resistance mutations and biochemical data. The inhibitor was found to bind in a pocket, on the upper part of the thumb domain, occupied in the apo-enzyme by a small helix, α-helix A, at the tip of the fingertip Λ1 loop (Figure 5.2). The binding pocket is located on the opposite side to the external GTP binding site respect to Pro495. However, whereas GTP interacts with amino acids in a α-helix A stabilizing its interaction with the thumb domain, binding of NNIs causes the displacement of the same helix thus freezing the enzyme in an inactive conformation, explaining the opposite effects *in vitro* on the activity of the enzyme activator, compared to site-A inhibitors.

5.1.2.2.   NNIs binding to site B

In addition to the NNI-binding site just described at the top of the thumb domain, several structurally diverse NNIs have been shown to bind into a 30 Å long hydrophobic cleft present near the base of the thumb domain.[195] This site, that here is referred as site B (Figure 5.2), is formed by the other side of the helices forming the floor of the site A binding pocket and the initial part of the polymerase C-terminal tail. Site B is 35 Å away from the active site and 10-15 Å from both the allosteric GTP binding site and from site A. Even though most of the amino acids composing this site are conserved across different genotypes and the only polymorphisms present in different viral isolates are found in the rim of the cleft, the polymerases form HCV genotypes different from 1 are not sensitive to inhibitors binding to this site.[206]

---

[205] Di Marco, S.; Volpari, C.; Tomei, L.; Altamura, S.; Harper, S.; Narjes, F.; Koch, U et al. Interdomain communication in hepatitis C virus polymerase abolished by small molecule inhibitors bound to a novel allosteric site. *J. Biol. Chem.* **2005**, *280*, 39260-39267.

[206] Ludmerer, S. W.; Graham, D. J.; Boots, E.; Murray, E. M; Simcoe, A, et al. Replication fitness and NS5B drug sensitivity of diverse hepatitis C virus isolates characterized by using a transient replication assay. *Antimicrob. Agents Chemother.* **2005**, *49*, 2059-2069.

NNIs with very different scaffolds have been shown to bind site B. The first allosteric compounds binding to this site were based on a *N,N*-disubstituted phenylalanine scaffold (Chart 5.3). These compounds, although active against the enzyme, were inactive in the replicon and therefore of limited interest.[207] The diphenyl alanine derivatives were followed by inhibitors based on 5-phenyl-thiophene scaffold (Chart 5.3) and a related carboxamide series, both with improved potency in the replicon assay.[208,209] Other inhibitors binding to site B were subsequently described such as dihydropirone analogues (Chart 5.3), pyranoindoles, carbazoles and cyclopentaindoles.[210,211] In a phase I study a pyranoindole developed by Viropharma/Wyeth (HCV-371; Chart 5.3) showed acceptable safety and tolerability but failed to show antiviral effects. Development of this series was therefore discontinued.

Crystal structures of the HCV polymerases bound to different site B inhibitors have been reported.[200,201,212] The inhibitors typically bind in the central region of the cleft, filling a dimple on the surface of the cleft formed by the side chains of Leu419, Arg422, Met423, Tyr477 and Trp528. In agreement with the structural finding mutations in the HCV-1 polymerase of either Met423 (M423L/V), Leu419 (L419M) are able to confer resistance to a variety of site B inhibitors (Figure 5.2).

---

[207] Chan, L.; Reddy, T. J.; Proulx, M.; Das, S. K.; Pereira, O.; Wang, W.; et al. Identification of *N,N*-disubstituted phenylalanines as a novel class of inhibitors of hepatitis C NS5B polymerase. *J. Med. Chem.* **2003**, *46*, 1283-1285.

[208] Chan, L.; Das, S. K.; Reddy, T. J.; Poisson, C.; Proulx, M.; Pereira, O. et al. Discovery of thiophene-2-carboxylic acids as potent inhibitors of HCV NS5B polymerase and HCV subgenomic RNA replication. Part 1: Sulfonamides. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 793-796.

[209] Chan, L.; Pereira, O.; Reddy, T. J.; Das, S. K.; Poisson, C. et al. Discovery of thiophene-2-carboxylic acids as potent inhibitors of HCV NS5B polymerase and HCV subgenomic RNA replication. Part 2: tertiary amides. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 797-800.

[210] Li, H.; Tatlock, J.; Linton, A.; Gonzalez, J.; Borchardt, A. et al. Identification and structure-based optimization of novel dihydropyrones as potent HCV NS5B polymerase inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4834-4838.

[211] Gopalsamy, A, Lim, K. et al. Discovery of pyrano[3,4-b]indoles as potent and selective HCV NS5B polymerase inhibitors. *J. Med. Chem.* **2004**, *47*, 6603-6608.

[212] Biswal, B. K.; Cherney, M.; Wang, M.; Chan, L.; Yannopoulos, C. G.; Bilimoria, D.; Nicolas, O.; Bedard, J.; James, M. N. Crystal structures of the RNA dependant RNA polymerase genotype 2a of HCV reveal two conformations and suggest mechanisms of inhibition by non-nucleoside inhibitors. *J. Biol. Chem.* **2005**, *280*, 18202-18210.
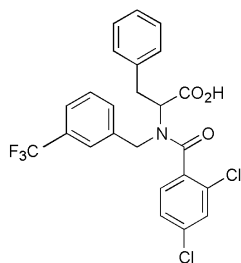
Phenylalanine derivative
Enzyme IC_{50} 0.7 μM
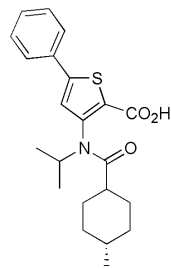
Thiophene 2-carboxilic acid derivative
Enzyme IC_{50} 1.5 μM
Replicon IC_{50} 0.6 μM
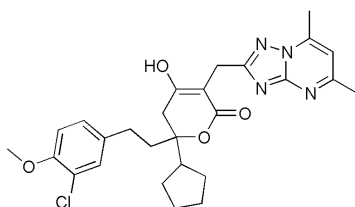
Dihydropyrone derivative
(AG-021541)
Enzyme IC_{50} 0.02 μM
Replicon IC_{50} 3.2 μM
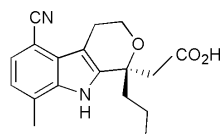
Piranoindole derivative
(HCV-371)
Enzyme IC_{50} 0.33 μM
Replicon IC_{50} 4.8 μM

Chart 5.3. Chemical structures of some selected site-B non-nucleoside inhibitors of the NS5B RdRp

Compared to site A inhibitors, the mechanism of inhibition of site B molecules is less well understood. In all, the currently available structural data[213,214,215] together with the observation that sites A and B share two α-helices suggests that inhibitors binding to these two

---

[213] Biswal, B. K.; Wang, M.; Cherney, M. M.; Chan, L.; Yannopoulos, C. G.; Bilimoria, D.; Bedard, J.; James, M. N. Non-nucleoside inhibitors binding to hepatitis C virus NS5B polymerase reveal a novel mechanism of inhibition. *J. Biol. Chem.* **2006**, *361*, 33-45.

[214] Howe, A. Y.; Cheng, H.; Thompson, I.; Chunduru, S. K.; Herrmann, S.; O'Connell, J. et al. Molecular mechanism of a thumb domain hepatitis C virus non-nucleoside RNA-dependent RNA polymerase inhibitor. *Antimicrob. Agents Chemother.* **2006**, *50*, 4103-4113.

[215] Wang, M.; Ng, K. K.; Cherney, M. M.; Chan, L.; Yannopoulos, C. G.; Bedard, J.; Morin, N. et al. Non-nucleoside analogue inhibitors bind to an allosteric site on HCV NS5B polymerase. Crystal structure and mechanism of inhibition. *J. Biol. Chem.* **2003**, *278*, 9489-9495.

sites may inhibit by using a similar mechanism, namely by interfering with the interaction between fingertip and thumb domains.[195]

5.1.2.3.   NNIs binding to sites C and D

Sites C and D are two overlapping sites located in the core of the polymerase in proximity of the enzyme active site cavity and at the interface between thumb and palm domains (Figure 5.2). Site C is located in a pocket in the active site cavity near the so-called primer grip, the C-terminal tail, the *β*-flap and the active site GDD motif. Several classes of compounds (Chart 5.4) are known to bind to this site,[201] including benzathiadiazines, acrylic acids derivatives, pyrrolidine acids, rhodanine and isothiazoles derivatives. Crystal structures of the genotype 1b HCV polymerase with several inhibitors bound to site C have been reported.[216,217,218,219] These structures show that the inhibitors are located in proximity of Met414 (Figure 5.2), which is consistent with the resistance data. Of note, for the rhodanines derivatives and isothiazoles the structure showed the formation of a covalent disulfide bond between the inhibitor and the sulphur of the strictly conserved Cys366.[220]

[216] Tedesco, R.; Shaw, A.N.; Bambal, R.; Chai, D.; Concha, N.O.; Darcy, M.G. et al. 3-(1,1-dioxo-2H-(1,2,4)-benzothiadiazin-3-yl)-4-hydroxy-2(1H)-quinolinones, potent inhibitors of hepatitis C virus RNA-dependent RNA polymerase. *J. Med. Chem.* **2006**, *49*, 971-983.

[217] Pfefferkorn, J.A.; Greene, M.L.; Nugent, R.A.; Gross, R.J.; Mitchell, M.A.; Finzel, B.C. et al. Inhibitors of HCV NS5B polymerase. Part 1: evaluation of the southern region of (2Z)-2-(benzoylamino)-3-(5-phenyl-2-furyl)acrylic acid. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2481-2486.

[218] Pfefferkorn, J.A.; Nugent, R.; Gross, R.J.; Greene, M.; Mitchell, M. A.; Reding, M. T. et al. Inhibitors of HCV NS5B polymerase. Part 2: evaluation of the northern region of (2Z)-2-benzoylamino-3-(4-phenoxy-phenyl)-acrylic acid. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2812-2818.

[219] Yan, S.; Appleby, T.; Gunic, E.; Shim, J.H.; Tasu, T.; Kim, H.; Rong, F.; Chen, H. et al. Isothiazoles as active-site inhibitors of HCV NS5B polymerase. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 28-33.

[220] Lee, G.; Piper, D.E.; Wang, Z.; Anzola, J.; Powers, J.; Walker, N.; Li, N. Novel inhibitors of hepatitis C virus RNA-dependent RNA polymerases. *J. Mol. Biol.* **2006**, *357*, 1051-1057.
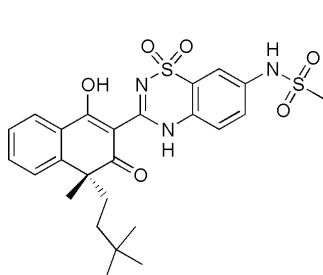
Thiadiazine derivative (A-848837)
Enzyme IC_{50} 0.3 - 0.6 nM
Replicon IC_{50} 2 nM
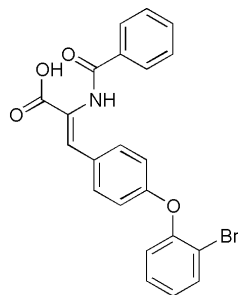
Acrylic acids derivative
Enzyme IC_{50} 100 nM
Replicon IC_{50} 28 nM
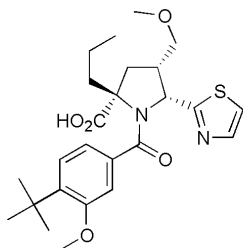
Pyrrolidine acid derivative
Enzyme IC_{50} 190 nM
Replicon IC_{50} 500 nM
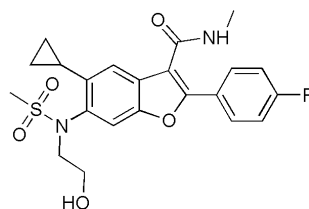
Benzofuran derivative (HCV-796)
Enzyme IC_{50} 0.16 nM
Replicon IC_{50} 8.6 nM

Chart 5.4. Chemical structures of some selected site-C and site-D non-nucleoside inhibitors of the NS5B RdRp

Inhibitors with a benzofuran scaffold (Chart 5.4) have been recently reported and the site of inhibition, site D, has been recently identified by X-ray crystallography.[221] These compounds, also referred as "hinge inhibitors", bind in the polymerase palm domain (Figure 5.2) and induce a hinge movement of the active site Arg200 side-chain to generate a relatively small pocket where the inhibitor is accommodated. One of these molecules, HCV-796, has gone

---

[221] Thommes, P.; Convery, M.; Williams, E.; Mesogiti, D.; Slater, M.; Bains, S.; Gray, F.; Haigh, D. The R200-hinge site, a novel inhibitor binding site on the HCV NS5B RNA-dependent RNA polymerase, *13th International Meeting on Hepatitis C Virus and Related Viruses, Cairns, Queensland, Australia, August 27-31,* **2006**, vol. *426*, Abstract.

through a Phase 1b proof-of-concept study.[222] HCV-796 inhibits RNA synthesis in both cell-free and replicon assays. Changes in amino acids adjacent to the active site conferred reduced susceptibility to HCV-796, with a major resistance mutation generated in the replicon being C316Y in phase 1b studies, the combination of HCV-796 with Peg-IFN and ribavirin resulted in an enhanced antiviral activity and an ~3.3-log-unit reduction in HCV RNA levels in infected patients. The clinical development of HCV-796 was, however, halted due to elevations in liver enzyme levels in the blood of some patients after prolonged treatment. Efforts to eliminate the side effects in this class of inhibitors may yet yield clinically useful compound.[223]

---

[222] Villano, S. A.; Howe, A. Y.; Raible, D.; Harper, D.; Speth, J.; Bichier, G. Analysis of HCV NS5B genetic variants following monotherapy with HCV-796, a non-nucleoside polymerase inhibitor, in treatment-naïve HCV infected patients. *Hepatology*, **2006**, *44*, 607A-608A.

[223] Flint, M; Mullen, S.; Deatly, A. M.; Chen, W.; Miller, L. Z.; Ralston, R.; Broom, C.; Emini, E. A.; Howe, A. Y. Selection and characterization of hepatitis C virus replicons dually resistant to the polymerase and protease inhibitors HCV-796 and boceprevir (SCH 503034). *Antimicrob. Agents Chemother*. **2009**, *53*, 401-11.

6. NS5B RESULTS ANALYSIS

Herein, will be described how the combination of a complete computational procedure together with biological studies led to the identification of novel molecular scaffolds, hitherto untested toward NS5B polymerase.[224]

Structure based 3D-QSAR models were generated employing NS5B non-nucleoside inhibitors (NNIs), whose bound conformations were readily available from the Protein Data Bank. These were grouped into two training sets of structurally diverse NS5B NNIs, based on their binding to the enzyme thumb (15 NNIs) or palm (10 NNIs) domains.

Ligand based (LB) and structure based (SB) alignments were rigorously investigated to assess the reliability on the correct molecular alignment for unknown binding mode modeled compounds. Both Surflex and Autodock programs were able to reproduce with minimal errors the experimental binding conformations of 24 experimental NS5B allosteric inhibitors. Eighty-one (thumb) and 223 (palm) modeled compounds taken from literature were LB and SB aligned and used as external validation sets for the development of 3D-QSAR models. Low error of prediction proved the 3D-QSARs to be useful scoring functions for the *in silico* screening procedure.

Finally, the virtual screening of the NCI Diversity Set led to the selection for enzymatic assays of 20 top-scoring molecules for each final model. Among the tested molecules, 10% resulted in inhibiting the NS5B RdRp activity at micromolar levels.

Binding mode analysis of hit compounds within the NS5B polymerase thumb domain showed that one of them, **NSC 123526**, exhibited a docked conformation which was in good agreement with the thumb training set most active compound (**6**).

---

[224] Musmuca, I.; Caroli, A.; Mai, A.; Kaushik-Basu, N.; Arora, P.; Ragno, R. Combining 3-D QSAR with Ligand Based and Structure Based Alignment Procedures for in Silico Screening of New Hepatitis C Virus NS5B Polymerase Inhibitors. *J. Chem. Inf. Model.* **2010**, *50*, 662-676

6.1. Ligand-Based, Structure-Based and 3D-QSAR Protocol

In any 3D-QSAR model preparation, after a careful selection of the training set, several steps need to be accomplished.[68,70,71,225] Among them, alignment rules (ligand-based or structure-based), statistical analysis (PLS and cross-validation)[52] and external validation (test sets) are of fundamental importance. Once these steps have been assessed, the model can be applied to predict the activity of untested molecules. The following section describes how we have addressed these steps, in order to develop an adequate procedure to efficiently predict the activity of new NS5B NNIs.

Training Sets **Selection**. At the beginning of these studies, 27 NS5B/NNIs complexes were available in the Protein Data Bank.[226] Analysis of the downloaded crystal structures revealed that ligands could be classified into three groups on the basis of their different binding sites on the NS5B RdRp (Figure 6.1). The first class consists of 15 crystallized NNIs that are believed to occupy a common binding site in the thumb subdomain (~35 Å from the polymerase active site) and thus, the ligands were collected and used as thumb training set.
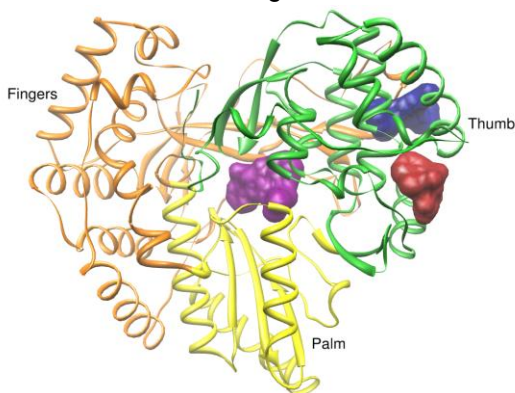
Figure 6.1. Ribbon show of the overall structure of NS5B RdRp with domains colored according to thumb (green), palm (yellow), and fingers (orange). Three allosteric binding sites surfaces are also shown. Dark red colored surface corresponds to the thumb allosteric bs, dark violet colored surface corresponds to the palm allosteric bs and the dark blue colored surface corresponds to the allosteric binding site situated in the thumb domain, near but clearly distinct from the first one (dark red surface).

---

[225] Clark, M.; Cramer, R. D., III; Jones, D. M.; Patterson, D. E.; Simeroth, P. E. Comparative molecular field analysis (CoMFA). 2. Toward its use with 3D-structural databases. *Tetrahedron Comput. Methodol.* **1990**, *3*, 47-59.
[226] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.

This set of compounds is structurally related to *N,N*-disubstituted phenylalanine,[207,215] thiophene,[208,209] thiazolone,[227,228,229] dihydropyrones,[210,230] and acyl pyrrolidine[231,232] molecular scaffolds. PDB codes together with ligand names, structures and biological activities of thumb training set compounds are represented in Table 6.1. The second class consisting of 10 crystallized NNIs, appear to bind approximately 10 Å from the catalytic site in the palm domain. Members of this class comprise the following NNIs: benzothiadiazine analogues,[216,233] acyl pyrrolidine analogues,[231,232] benzylidene analogues,[220,234] acrylic acid derivatives,[217,218] and proline sulfonamide analogues.[235] These ligands were collected and used as palm training set and their PDB codes, structures and

[227] Yan, S.; Appleby, T.; Larson, G.; Wu, J. Z.; Hamatake, R.; Hong, Z.; Yao, N. Structure-based design of a novel thiazolone scaffold as HCV NS5B polymerase allosteric inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5888-5891.

[228] Yan, S.; Larson, G.; Wu, J. Z.; Appleby, T.; Ding, Y.; Hamatake, R.; Hong, Z.; Yao, N. Novel thiazolones as HCV NS5B polymerase allosteric inhibitors: Further designs, SAR, and X-ray complex structure. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 63-67.

[229] Yan, S.; Appleby, T.; Larson, G.; Wu, J. Z.; Hamatake, R. K.; Hong, Z.; Yao, N. Thiazolone-acylsulfonamides as novel HCV NS5B polymerase allosteric inhibitors: convergence of structure-based drug design and X-ray crystallographic study. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1991-1995.

[230] Love, R. A.; Parge, H. E.; Yu, X.; Hickey, M. J.; Diehl, W.; Gao, J.; Wriggers, H.; Ekker, A.; Wang, L.; Thomson, J. A.; Dragovich, P. S.; Fuhrman, S. A. Crystallographic identification of a non-competitive inhibitor binding site on the hepatitis C virus NS5B RNA polymerase enzyme. *J. Virol.* **2003**, *77*, 7575-7581.

[231] Burton, G.; Ku, T. W.; Carr, T. J.; Kiesow, T.; Sarisky, R. T.; Lin-Goerke, J.; Baker, A.; Earnshaw, D. L.; Hofmann, G. A.; Keenan, R. M.; Dhanak, D. Identification of small molecule inhibitors of the hepatitis C virus RNA-dependent RNA polymerase from a pyrrolidine combinatorial mixture. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1553-1556.

[232] Slater, M. J.; Amphlett, E. M.; Andrews, D. M.; Bravi, G.; Burton, G.; Cheasty, A. G.; Corfield, J. A.; Ellis, M. R.; Fenwick, R. H.; Fernandes, S.; Guidetti, R.; Haigh, D.; Hartley, C. D.; Howes, P. D.; Jackson, D. L.; Jarvest, R. L.; Lovegrove, V. L.; Medhurst, K. J.; Parry, N. R.; Price, H.; Shah, P.; Singh, O. M.; Stocker, R.; Thommes, P.; Wilkinson, C.; Wonacott, A. Optimization of novel acyl pyrrolidine inhibitors of hepatitis C virus RNA-dependent RNA polymerase leading to a development candidate. *J. Med. Chem.* **2007**, *50*, 897-900.
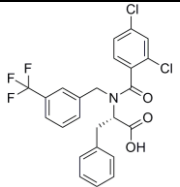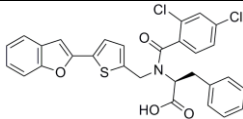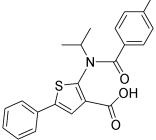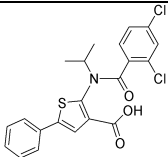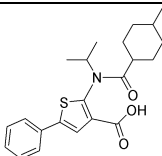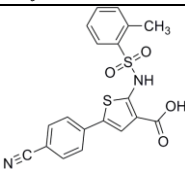
[233] Dhanak, D.; Duffy, K. J.; Johnston, V. K.; Lin-Goerke, J.; Darcy, M.; Shaw, A. N. et al. Identification and biological characterization of heterocyclic inhibitors of the hepatitis C virus RNA-dependent RNA polymerase. *J. Biol. Chem.* **2002**, *277*, 38322-38327.
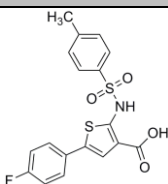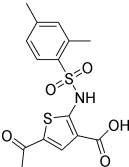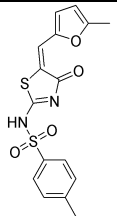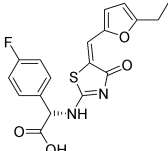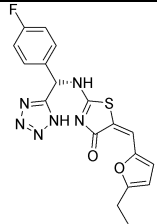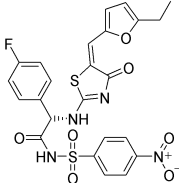
[234] Powers, J. P.; Piper, D. E.; Li, Y.; Mayorga, V.; Anzola, J.; Chen, J. M. et al. SAR and mode of action of novel non-nucleoside inhibitors of hepatitis C NS5b RNA polymerase. *J. Med. Chem.* **2006**, *49*, 1034-1046.
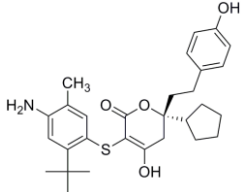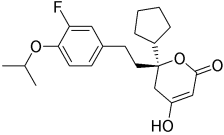
[235] Gopalsamy, A.; Chopra, R.; Lim, K.; Ciszewski, G. et al. Discovery of proline sulfonamides as potent and selective hepatitis C virus NS5b polymerase inhibitors. Evidence for a new NS5b polymerase binding site. *J. Med. Chem.* **2006**, *49*, 3052-3055.

biological activities are represented in Table 6.2. It may be noted that compound **15**, an acyl pyrrolidine derivative, (PDB entry code: 2JC0),[231,232] has been included in both training sets, since the available co-crystallized structure of this compound showed it bound in both allosteric pockets.

Table 6.1. PDB Code, Ligand Name and Ligand Structures of the 15 NS5B polymerase NNIs that represent the thumb training set.

PDB Code	Ligand Name	Entry	Ligand Structure	IC$_{50}$ (μM)
1NHU	(2S)-2-[(2,4-dichloro-benzoyl)-(3-trifluoromethyl-benzyl)-amino] -3-phenyl-propionic acid	1		1.7
1NHV	(2S)-2-[(5-benzofuran-2-yl-thiophen-2-yl methyl)-(2,4-dichloro-benzoyl)-amino]-3-phenyl-propionic acid	2		8.6
1YVX	3-[isopropyl(4-methyl benzoyl) amino]-5-phenylthiophene-2-carboxylic acid	3		8.0
1YVZ	3-[(2,4-dichlorobenzoyl) (isopropyl)amino]-5-phenyl thiophene-2-carboxylic acid	4		4.4
2GIR	3-{isopropyl[(trans-4-methyl cyclohexyl)carbonyl]amino}-5-phenylthiophene-2-carboxylic acid	5		1.5
2D3U	5-(4-cyanophenyl)-3-{[(2-methylphenyl)sulfonyl] amino}thiophene-2-carboxylic acid	6		0.2663
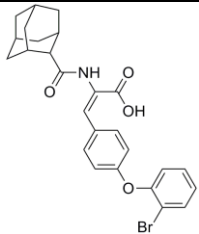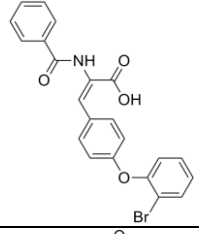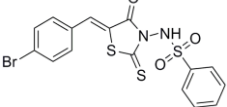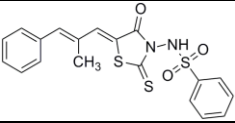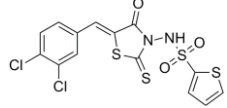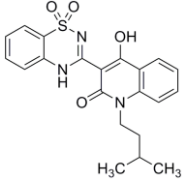
PDB Code	Ligand Name	Entry	Ligand Structure	IC$_{50}$ (μM)
2D3Z	5-(4-fluorophenyl)-3-{[(4-methylphenyl) sulfonyl] amino}thiophene-2-carboxylic acid	**7**		0.2918
2D41	5'-acetyl-4-{[(2,4-dimethyl phenyl) sulfonyl]amino} - 2,2'-bithiophene-5-carboxylic acid	**8**		0.3073
2HWH	4-methyl-N-{(5E)-5-[(5-methyl-2-furyl)methylene]-4-oxo-4,5-dihydro-1,3-thiazol-2-yl} benzenesulfonamide	**9**		2.0
2HWI	(2S)-({(5Z)-5-[(5-ethyl-2-furyl)methylene]-4-oxo-4,5-dihydro-1,3-thiazol-2-yl}amino)(4-fluorophenyl) acetic acid	**10**		3.0
2I1R	(5Z)-5-[(5-ethyl-2-furyl) methylene]-2-{[(S)-(4-fluoro- phenyl)(1H-tetrazol-5-yl) methyl]amino} -1,3-thiazol-4(5H)-one	**11**		9.7
2O5D	(2S)-2-({(5Z)-5-[(5-ethyl-2-furyl)methylene]- 4-oxo-4,5-dihydro-1,3-thiazol-2-yl} amino)-2-(4-fluorophenyl)-N-[(4-nitrophenyl)sulfonyl] acetamide	**12**		7.0

PDB Code	Ligand Name	Entry	Ligand Structure	IC$_{50}$ (μM)
1OS5	3-(4-amino-2-tert-butyl-5-methyl-phenylsulfanyl)- 6-cyclopentyl-4-hydroxy-6-[2-(4-hydroxy- phenyl)-ethyl]-5,6-dihydro-pyran-2-one	**13**		0.93
2HAI	(6S)-6-cyclopentyl-6-[2-(3-fluoro-4-isopropoxyphenyl) ethyl]- 4-hydroxy-5,6-dihydro-2H-pyran-2-one	**14**		0.53
2JC0	(2S,4S,5R)-2-isobutyl-5-(2-thienyl)-1-[4-(trifluoromethyl) benzoyl]pyrrolidine-2,4-dicarboxylic acid	**15**		20.0

A third class of compounds consists of indole-derived NNIs[205,236] that was reported to bind to an allosteric pocket in the thumb domain, close in space but clearly distinct (Figure 6.1) from the cleft occupied by the above described thumb training set. In Table 6.3 are shown PDB codes, ligand names, structures and biological activities of the third class. However, due to the limited number (three) of crystallized ligands in this group, it was not possible to use them as a further training set. Thus, in this study a total of 24 NNIs/NS5B complexes were utilized, representing 15 thumb and 10 palm allosteric NNIs, with 2JC0 complex included in both categories. All subsequent 3D-QSAR studies, molecular docking simulations and ligand-based alignments were carried with these two training sets.

---

[236] Ikegashira, K.; Oka, T.; Hirashima, S.; Noji, S.; Yamanaka, H. Discovery of conformationally constrained tetracyclic compounds as potent hepatitis C virus NS5B RNA polymerase inhibitors. *J. Med. Chem.* **2006**, *49*, 6950-6953.

Table 6.2. PDB Code, Ligand Name and Ligand Structures of the 10 NS5B polymerase NNIs that represent the palm training set.

PDB Code	Ligand Name	Entry	Ligand Structure	IC$_{50}$ (μM)
1Z4U	(2Z)-2-[(1-adamantylcarbonyl) amino]-3-[4-(2-bromophenoxy) phenyl]prop-2-enoic acid	**16**		0.07
1YVF	(2Z)-2-(benzoylamino)-3-[4-(2-bromophenoxy)phenyl]- 2-propenoic acid	**17**		0.10
2AWZ	(5R)-(4-bromophenylmethyl)-3-(benzenesulfonylamino)- 4-oxo-2-thionothiazolidine	**18**		0.70
2AX0	(5R)-(2E-methyl-3-phenyl-allyl)-3-(benzene sulfonylamino)-4-oxo-2-thionothiazolidine	**19**		0.20
2AX1	(5R)-(3,4-dichlorophenylmethyl)-3-(2-thiophenesulfonylamino)-4-oxo-2-thionothiazolidine	**20**		0.20
2FVC	3-(1,1-dioxido-4H-1,2,4-benzo-thiadiazin-3-yl)--4-hydroxy -1-(3-methyl-butyl)quinolin-2(1H)-one	**21**		0.032

135
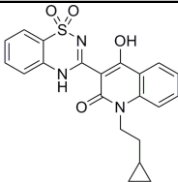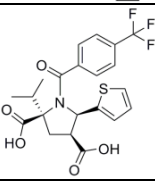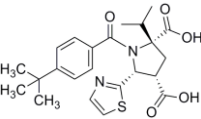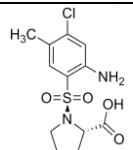
PDB Code	Ligand Name	Entry	Ligand Structure	IC50 (μM)
2GIQ	1-(2-cyclopropylethyl)-3-(1,1-dioxido-2H -1,2,4-benzothia-diazin-3-yl)-6-fluoro-4-hydroxy-quinolin-2(1H)-one	22		0.010
2JC0	(2S,4S,5R)-2-isobutyl-5-(2-thienyl)-1-[4 - (trifluoro methyl) benzoyl]pyrrolidine-2,4-dicarboxylic acid	15		20.0
2JC1	(2S,4S,5R)-1-(4-tert-butylbenzoyl)-2-isobutyl- 5-(1,3-thiazol-2-yl)pyrrolidine-2,4-dicarboxylic acid	23		3.80
2GC8	1-[(2-amino-4-chloro-5-methyl phenyl)sulfonyl]- L-proline	24		3.10

Table 6.3. PDB Code, Ligand Name and Ligand Structures of the indole-based NNIs.

PDB Code	Ligand Name	Ligand Structure	IC$_{50}$ (μM)
2BRK	3-cyclohexyl-1-(2-morpholin-4-yl-2-oxoethyl)-2-phenyl-1H-indole-6-carboxylic acid		0.026
2BRL	3-cyclohexyl-1-(2-{methyl[(1-methyl piperidin-3-yl) methyl] amino}-2-oxoethyl)-2-phenyl-1H-indole-6-carboxylic acid		0.018
2DXS	N-[(13-cyclohexyl-6,7-dihydro-indolo[1,2-[1,4]benzooxazepin -10-yl)carbonyl]-2-methyl-L-alanine		0.046

Polymerase-Inhibitor Complex Structures Preparation. The NS5B RNA polymerase structures co-crystallized with allosteric inhibitors were downloaded from the Brookhaven Protein Data Bank. Of the 24 complexes, 15 were co-crystallized with ligands that bind the thumb subdomain of NS5B. The remaining 10 complexes were co-crystallized with palm binding site NNIs. In order to reduce the geometric inaccuracy of the downloaded structures, the complexes were submitted to a molecular modeling protocol. All complexes were arbitrary superimposed using as template 2D3U, because it was one of the better resolved (R = 2.0 Å) and with the smallest number of missing residues. The superimposition of the NS5B complexes was carried out by means of the program Chimera[172] using the command-line implementation

of MatchMaker.[237] All crystal waters were discarded, and hydrogen atoms were added using the tleap module of AMBER suite.[238,239] All ligands were individually inspected, and the correct protonation states at pH 7.4 were considered, i.e., lysines, arginines, aspartates, and glutamates were assumed to be in the ionized form and the parameter were calculated by means of the Antechamber module of AMBER suite. The complexes were solvated (SOLVATEOCT command) in a box extending 10 Å with water molecules (TIP3 model) and neutralized with Na^+ and Cl^- ions. The solvated complexes were then refined by a single point minimization using the Sander module of AMBER suite. Once minimized, all complexes were realigned with MatchMaker using the same reference complex as template. Subsequently, the minimized ligands were extracted from the proper minimized complex (experimental alignments) and used as training sets for the generation of two structure-based statistical models (3D-QSARs).

Molecular Interaction Fields Calculation. The GRID force field was used to describe the previously superimposed molecular structures. Interaction energies between the selected probes and each molecule were implemented using a grid spacing of 1 Å. The *xyz* coordinates (in angstroms) of the grid rectangular box used for the computation are $X_{min}/X_{max} = -4.0/52,0$, $Y_{min}/Y_{max} = 6.0/78.0$, $Z_{min}/Z_{max} = 13.0/80.0$. Initially, many probes were tested (C1=, sp^2 carbon atom bonded to one hydrogen; C3, sp^3 aliphatic carbon atom bonded to three hydrogen atoms; OH, oxygen atom bearing one acidic hydrogen in phenols or carboxy –COOH; O, explicit resonating sp^2 oxygen with 2 lone pairs in carboxy acid anion or

---

[237] Meng, E. C.; Pettersen, E. F.; Couch, G. S.; Huang, C. C.; Ferrin, T. E. Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* **2006**, *7*, 339.

[238] Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1-41.

[239] Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668-1688.

pyridine oxide C_5H_5NO; N1, planar nitrogen bonded to one hydrogen atom and two other atoms, no lone pair; N1=, cationic sp^2 nitrogen with one hydrogen, one other single and one double bond; N3+, cationic sp^3 nitrogen bonded to 3 hydrogens and with one other rotatable bond; DRY, hydrophobic). Preliminary r^2, q^2 and external standard deviation errors of prediction (SDEP) values by PLS, cross-validation and external validations (see Tables S1, S2, S3 &S4 in the Supporting Information) suggested the use of C1= probe for final 3D-QSAR interpretations of both training sets. Attempts to use a combination of different GRID probes did not lead to any improvement of the statistical models.

Statistical Analyses. The MIFs of the training sets were imported into the GOLPE program. The associated biological activities expressed as IC_{50} (concentrations reflecting 50% inhibition of HCV NS5B RdRp activity) values (in μM), were converted to molar pIC_{50} (-log IC_{50}), and used as dependent variables in 3D-QSAR analysis. The correlation between biological activities and MIFs was carried out by a statistical PLS analysis as implemented in GOLPE program. Thus, a series of 3D-QSAR models for the thumb and the palm training sets were obtained (Tables S1 & S2 in the Supporting Information). The goodness of each model was measured by the statistical indices r^2, q^2 and SDEP using leave-some-out with five random groups (LSO-5) and cross-validation method.

The resulting probe−target interaction energies for each compound were first unfolded to produce one-dimensional vector variables and then assembled in the so-called X matrix. The matrix was pretreated using a cutoff of 5 kcal/mol, in order to produce a more symmetrical distribution of energy values. Variable pre-selection was operated by zeroing values with absolute values smaller than 0.01 kcal/mol and removing variables with standard deviation below 0.05. In addition, variables taking only two and three distributions were also removed. Attempts to use other settings in the variable pre-selection did not lead to any improvement of the models.

The smart region definition (SRD) algorithm was applied. This is a GOLPE implementation aimed at selecting cluster of variables, rather than the single variable mainly responsible for activity. The SRD technique seems less prone to change correlation than any single variable selection, and improves the interpretability of the models. A number of seeds (1000) were selected using a D-optimal design criterion in the weight space. Structural differences between different molecules in the series will be reflected in groups of variables and, therefore, groups were generated around each seed in the 3D space. Variables with a critical distance cutoff of 2 Å to the seeds were included in the groups. If two neighboring groups (with a distance smaller than 10 Å) contained the same information, the groups were collapsed. The groups were used in the variable selection procedure replacing the original variables. The effect of the groups on the predictability was evaluated and, groups instead of the individual variables were removed from the data file. Attempts to use other settings in the SRD did not lead to any improvement of the models.

All 3D-QSARs were internally tested by cross-validation using LSO-5 method with 100 randomizations. The optimum number of principal components used in any 3D-QSAR model was chosen on the basis of the different cross-validated PLS analyses, setting to 5 the maximum number of principal components to be extracted. In the present study, the optimal number of principal components refers to the ones where $SDEP_{CV}$ assumes the minimum value and the cross-validation correlation coefficient q^2 assumes the maximum value. Attempts to increase the maximum principal components during CVs did not lead to any substantially different model.

Models Interpretations. The GRID/GOLPE method used within this study led to the definition of two final Structure–Based 3D-QSAR models. The statistical results obtained by using the C1= GRID probe are reported in Table 6.4. The internal predictive power of each model was checked by the leave-some-out with five random groups (LSO-5) cross-validation method (Figure 6.2). The

latter indicated that the optimal principal component was the third one. In fact the 2nd, 4th and 5th principal components showed lower q^2 and higher SDEP values (data not shown).
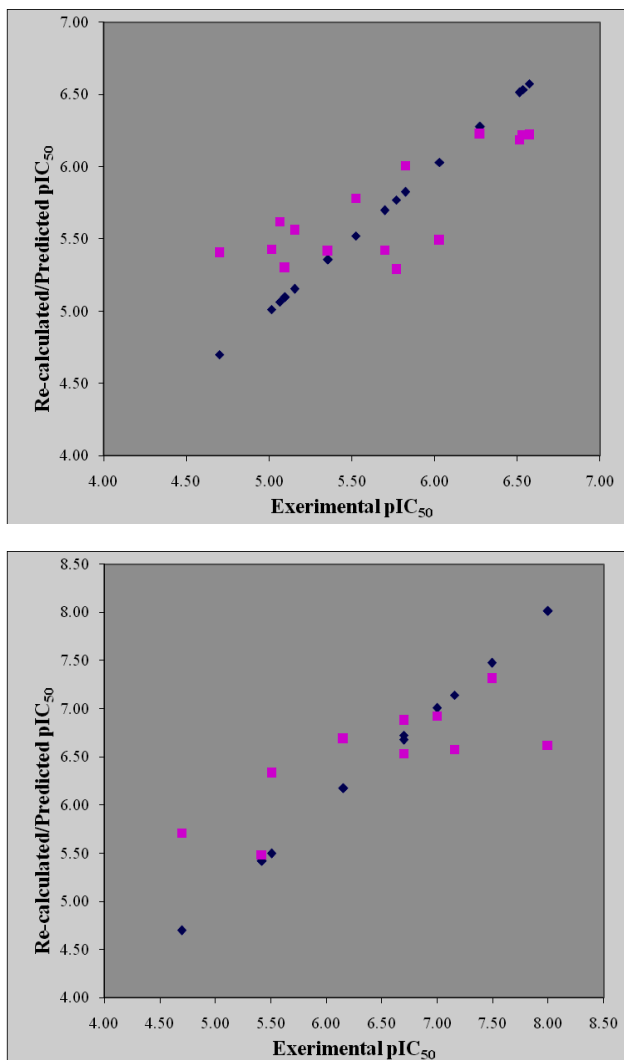
Figure 6.2. Fitting and Cross-Validation plots for the thumb (left) and palm (right) training sets.

As represented in Table 6.4, the statistical analysis yielded highly significant statistical coefficients with r^2 values of 0.99 for both models and q^2 of 0.69 for the thumb model and of 0.55 for the palm model, thus providing the goodness of the system.

Table 6.4. PLS Analysis Results for the Thumb and the Palm Structure-Based 3D-QSAR Models.[a]

model	N	GRID Probe	V	PC	r^2	q^2	SDEP$_{CV-L5O}$
Thumb TS	15	C1=	5133	3	0.99	0.69	0.31
Palm TS	10	C1=	3848	3	0.99	0.55	0.66

[a] N, number of compounds in the training set; V, number of GOLPE variables; PC, optimal number of principal components; r^2, conventional square correlation coefficient; q^2, cross-validation correlation coefficient; SDEP, cross-validated standard error of prediction using the leave-five-out cross-validation method

Moreover, an accurate graphical analysis of GOLPE contour maps led to the identification of the most representative areas selected by the models. In fact, one of the most interesting features of a GRID/GOLPE 3D-QSAR analysis is the possibility of translating back the PLS coefficients assigned to each variable to the 3D positions they occupy in real space. These values can be contoured at a particular significant level and can be displayed as a grid plot of PLS coefficients. The contour coefficient maps indicate those areas in which the model has found a high correlation between the ligand-probe interaction energy and the biological activity. However, the signs of PLS coefficients can induce errors in this plot. Coefficients could have opposite meaning depending on the positive or negative field values produced by the compound in the same area. In Figure 6.3 are shown the PLS coefficient plots for the thumb training set analyzed with the C1= GRID probe. One of the most important areas selected by the model is represented by a big red polyhedron located between the *p*-cyanophenyl thiophene and the 2-methylphenyl moieties of the most active compound (**6**), corresponding to the thiazole and the 4-F phenyl group of one of the lowest active compounds (**11**). For chemical interpretation, one must remember that red contours in coefficient maps suggest that repulsive interactions (positive GRID energies)

increase affinity, while attractive interactions (negative GRID energies) decrease activity. The sp^2 aromatic carbon atom used for the calculation of the molecular field is a prevalently hydrophobic group. Thus, it might be suggested that these areas corresponds to hydrophobic surface of the thumb domain where van der Waals interactions as well as $\pi-\pi$ interactions with binding site residues, positively contributes to the biological activity. On the other hand a blue colored region around the tetrazole portion of compound **11** clearly indicates that the steric hindrance in this area is not suitable for potency.
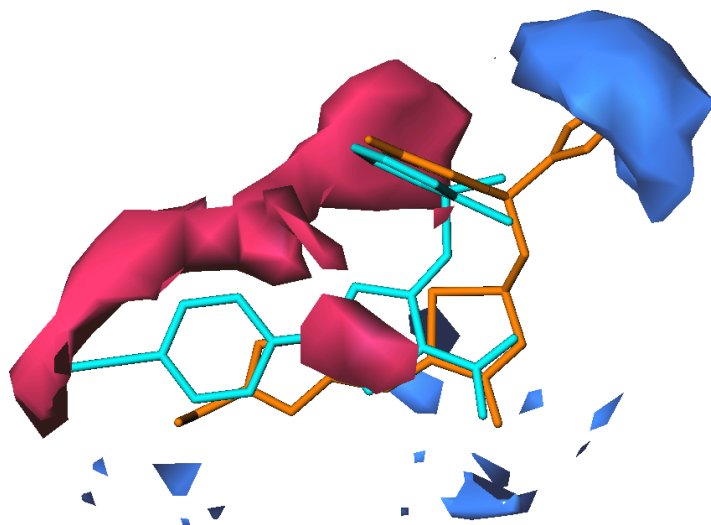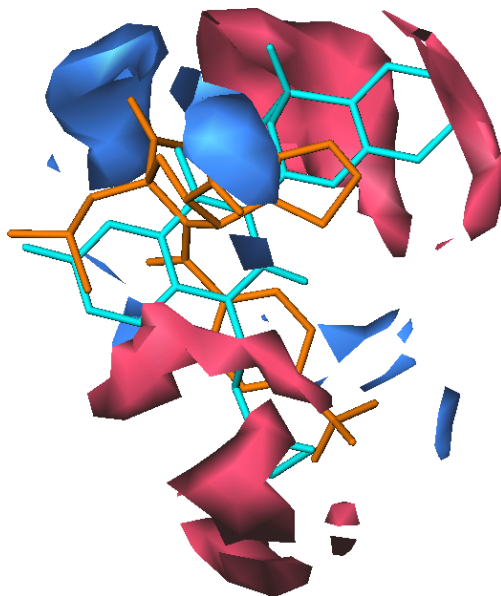

Figure 6.3. Contour maps of the PLS Coefficients derived from C1= GRID probe analysis using the 15 compounds of the thumb training set (contour levels: 0.0008 red, -0.0008 blue). To aid interpretation only the highest active (**6** in cyan) and one of the lowest active (**11** in orange) compounds are shown. Hydrogen atoms are omitted for the sake of clarity.

In Figure 6.4 some small blue polyhedra near the carboxylate group of compound **6** are also depicted. To shed light on the nature of the interactions made by this substituent, we decided to analyze it by means of OH GRID probe derived 3D-QSAR (for statistical data see Tables S1 & S2 in the Supporting Information).

The hydroxyl probe used for the calculation of the molecular field is a prevalently polar group with the ability to participate in hydrogen bonds and electrostatic interactions. Thus, areas containing negative coefficients (blue polyhedrons) correspond to areas in the binding pocket where energetically favorable interactions produce an increase in the activity. These regions correspond to position of polar groups in the binding site. Consequently, blue polyhedra in this area mainly reflect hydrophilic features, which may be due to the presence of hydrogen bond donor residues in the binding pocket.
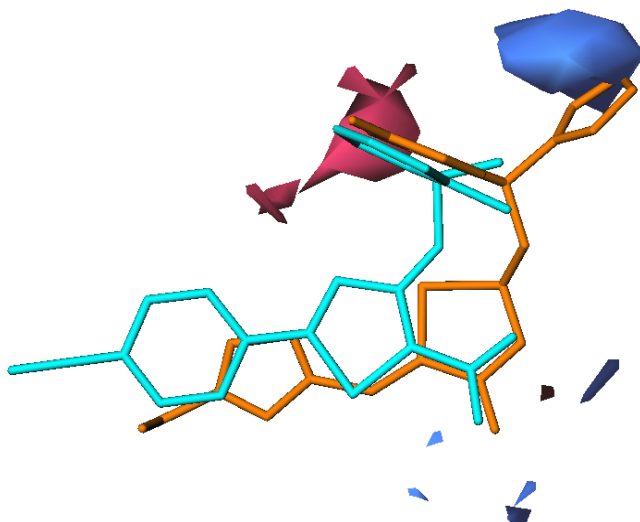
Figure 6.4. Contour maps of the PLS Coefficients derived from OH GRID probe using the 15 compounds of the thumb training set (contour levels: 0.0008 red, -0.0008 blue). To aid interpretation only the highest active (**6** in cyan) and one of the lowest active (**11** in orange) compounds are shown. Hydrogen atoms are omitted for the sake of clarity.

In Figure 6.5, the PLS coefficients plots for the palm training set are shown. The 3D-QSAR model graphical analyses was carried out by means of C1= probe. The red polyhedron situated around the benzo portion of the benzothiadiazine system of compound **22** appears to be relatively important for the biological activity. In fact,

as previously reported,[216] a key feature observed in all of the benzothiadiazine structures determined is an apparent edge-to-face π–interaction between Phe193 and the benzene ring. Compound **15**, an acyl pyrrolidine derivative, did not show any particular interaction in this area. This may be due to the relevant differences in the chemical structures of these molecules. Nevertheless, the red polyhedra located in proximity of the *N*-1 (2-cyclopropyl)ethyl moiety (**22**) and the 4-trifluoromethyl benzoyl group of compound **15** imply that space filling in both positions might play a key role in further potency enhancements.

Figure 6.5. Contour maps of the PLS Coefficients derived from C1= Grid probe using the 10 compounds of the palm training set (contour levels: 0.0010 red, -0.0010 blue). To aid interpretation only the highest active (**22** in cyan) and the lowest active (**15** in orange) are shown. Hydrogen atoms are omitted for the sake of clarity.

6.2. Alignment Rules

In the present study, ligand-based as well as structure-based alignment methodologies were explored with the aim to determine the alignment of molecules with unknown binding mode (test set). Several software that perform diverse alignment procedures are available, and among them, Surflex-Sim[164,165] and Autodock[23] were chosen since free for academics and of good quality as assessed by considerable references.[240,241,242,243,244,245,246] In particular, for the LB approach, the application of the concept of morphological similarity, implemented in Surflex-Sim was used, while for the SB approach the program Autodock was selected.

In the following paragraphs different levels of alignment assessment are reported for either LB or SB methods. Regarding LB approach, randomly built conformation for each ligand was generated and then aligned to the experimental binding conformation (re-surflex). This way the LB alignment procedure was first tested for self alignment and due to different topological atom descriptions between the experimental and modeled

[240] Cleves, A. E.; Jain, A. N. Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.* **2006**, *49*, 2921-2938.

[241] Holt, P. A.; Ragazzon, P.; Strekowski, L.; Chaires, J. B.; Trent, J. O. Discovery of novel triple helical DNA intercalators by an integrated virtual and actual screening platform. *Nucleic Acids Res.* **2009**, *37*, 1280-1287.

[242] Langham, J. J.; Cleves, A. E.; Spitzer, R.; Kirshner, D.; Jain, A. N. Physical binding pocket induction for affinity prediction. *J. Med. Chem.* **2009**, *52*, 6107-6125.

[243] Ragno, R.; Mai, A.; Massa, S.; Cerbara, I.; Valente, S.; Bottoni, P.; Scatena, R.; Jesacher, F.; Loidl, P.; Brosch, G. 3-(4-Aroyl-1-methyl-1H-pyrrol-2-yl)-N-hydroxy-2-propenamides as a new class of synthetic histone deacetylase inhibitors. 3. Discovery of novel lead compounds through structure-based drug design and docking studies. *J. Med. Chem.* **2004**, *47*, 1351-1359.

[244] Wang, M.; Zhang, J.; Andrei, D.; Kuczera, K.; Borchardt, R. T.; Wnuk, S. F. Are L-adenosine and its derivatives substrates for S-adenosyl-L-homocysteine hydrolase? *J. Med. Chem.* **2005**, *48*, 3649-3653.

[245] Taylor, C. M.; Barda, Y.; Kisselev, O. G.; Marshall, G. R. Modulating G-protein coupled receptor/G-protein signal transduction by small molecules suggested by virtual screening. *J. Med. Chem.* **2008**, *51*, 5297-5303.

[246] Radi, M.; Falciani, C.; Contemori, L.; Petricci, E.; Maga, G.; Samuele, A.; Zanoli, S.; Terrazas, M.; Castria, M.; Togninelli, A.; Este, J. A.; Clotet-Codina, I.; Armand-Ugon, M.; Botta, M. A multidisciplinary approach for the identification of novel HIV-1 non-nucleoside reverse transcriptase inhibitors: S-DABOCs and DAVPs. *ChemMedChem* **2008**, *3*, 573-593.

structures, the surflex-aligned conformation was retained as a reference conformation in the subsequent calculations involving the use of modeled structures. Although arbitrary, this assumption was supported by the visual inspection of superimposed experimental versus the re-surflexed conformations (Figure 6.6). In the next step, each randomly built conformation is aligned to a target list which contains all ligands (in their binding conformation), except the corresponding experimental one (cross-surflex).
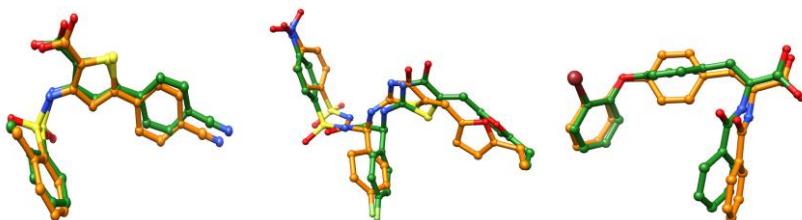
Figure 6.6. Examples of Surflex alignment. Superimposition of the modeled ligand conformations (carbon atoms in green) to the experimental ones (carbon atoms in orange) of three compounds of training sets (from left to right, compounds **6**, **12** and **16**). Atom bonds are in ball and stick fashion. Hydrogen atoms are omitted for the sake of clarity.

Regarding the SB approach, a four steps procedure was set up. In the first step, Autodock program was tested for the ability to reproduce the bound state of a certain ligand using the respective receptor and the known binding conformation as starting point (re-docking). In the second step, the same re-docking was repeated using as starting point a randomly generated conformation (re-docking modeled). The third level of docking assessment was achieved by a cross-docking procedure using the experimental conformation docked in all the available receptors while excluding the native one (cross-docking). In the last step the cross-docking procedure was repeated using the modeled ligands (cross-docking modeled). In either cross-docking procedures, the output conformations were merged in a single file and then clustered using the internal clustering procedure of the Autodock program. The rationale of the docking assessment resides in the fact that

first the docking program was tested for the ability to correctly reposition the native ligand conformation within the respective experimental receptor pocket. Besides, a modeled ligand conformation was used in order to measure how the starting docking conformation can influence the docking program in the reproduction of the bound ligand conformation. Analogously cross-docking experiments, using the experimental and modeled ligand conformations, were carried out to evaluate the docking program aptitude to predict the correct binding conformation using an ensemble of receptor conformations while keeping out the native co-crystallized one. In the cross-docking assessments, the clustering of poses coming from docking runs using different receptor conformations can be assumed a sort of full flexible docking simulation in which both the ligand and the receptor are allowed to move. In the case of the ligand the flexibility is achieved by the Solis-Wets algorithm,[247] while for the receptor the experimental NS5B snapshots give the induced ligand conformational changes. Attempt to use Autodock side-chain flexibility feature was not successful (data not shown).

6.2.1.   Ligand-Based Alignment: Surflex Assessment

The ligand-based alignment of the molecules was achieved using Surflex-Sim. This method optimizes the pose of a query molecule to an object molecule in order to maximize 3D similarity. Morphological similarity is defined as a Gaussian function of the differences in molecular surface distances of two molecules at weighted observation points on a uniform grid, thus yielding a value from 0 to 1. The function is dependent on the relative alignment of two molecules.

First step of Surflex-Sim assessment was checking the goodness of the alignment when a randomly built ligand conformation is generated and then overlapped to the experimental binding conformation (self alignment). For this purpose, each of the

---

[247] Solis, F. J.; Wets, R. J. B. Minimization by Random Search Techniques. *Math. Oper. Res.* **1981**, *6*, 19-30.

24 modeled compounds of the training sets were superimposed to the corresponding minimized ligands as extracted from their respective complex (re-surflex). Due to the differences in topological description of the molecules, the goodness of the self-alignment was visually inspected. As an example of the alignment goodness, in Figure 6.6 are depicted the aligned conformations of three different modeled compounds (**6**, **12** and **16)** superimposed on their respective experimental bound conformations.

In the next alignment assessment step, each of the 24 modeled ligand conformations was overlapped to a target list containing all the ligands (in their SB experimental aligned binding conformations), except the corresponding experimental one (cross-surflex). Here, the self-aligned surflex conformation was employed as a reference for measuring the root mean square deviation (rmsd) values between the cross-surflex aligned and the re-surflex one. As reported in Table S7 of the Supporting Information, the rmsd values for all 24 NS5B NNIs were below the cutoff value of 1, thus confirming a good alignment result. This validation of Surflex-Sim's ability to accurately align randomly built conformations suggested us that it would exhibit a similar accuracy even with molecules of unknown binding conformation. Therefore, it can be considered as a useful tool for LB alignment of NS5B inhibitors (see external test set below) as well as for the alignment of molecules in a virtual screening protocol (see below).

6.2.2.   Structure-Based Alignment: General Docking Settings

Autodock 4 was used for all docking calculations. The AutoDockTools (ADT) package was employed to generate the docking input files and analyze the docking results. The 24 ligands extracted from the minimized complexes (see Polymerase-Inhibitor Complex Structures Preparation section) were separated into two groups on the basis of their different binding sites on the NS5B polymerase receptor. The first group contains 15 NNIs that are believed to occupy a common binding site located in the thumb domain. The second group contains a total of 10 NNIs reported as

binding an allosteric pocket located near the active site in the palm domain.

Two different grid boxes, one for each binding site, were centered on the average mass center of the ligands. Thus, a grid box of 91 × 91 × 86 points and a grid spacing of 0.375 Å was set in order to accommodate the NNIs that bind the thumb domain of NS5B polymerase. The second grid box (spacing, 0.375 Å) of 62 × 75 × 75 points was implemented in such a way to accommodate the NNIs that bind the palm allosteric site of NS5B. Several atom probes (characterized by the same stereoelectronics properties as the atoms constituting the inhibitors) were moved on the grid nodes, while interaction energy between the probe and the inhibitors were calculated at each node. The, grid maps were generated for each atom probe, describing its interactions with the inhibitors. Autogrid 4, as implemented in the Autodock software package, was used to generate grid maps.

The Lamarckian genetic algorithm (LGA)[23] was employed to generate orientations or conformations of the ligands within the binding site. The global optimization started with a population of 150 randomly positioned individuals, a maximum of 2.5×10^6 energy evaluations, and a maximum of 27000 generations. A total of 100 runs were performed, while all the remaining run parameters were maintained at their default settings. A cluster analysis was carried out using 2 Å as the root-mean-square deviation tolerance. Docking experiments were also tried using a single grid box comprising the two binding sites. Although Autodock performed quite well in selectively positioning the ligands in the right pockets, the results in term of root-mean-square-deviation (rmsd) values were not fully satisfying (data not shown). This is likely due to the fact that with only 100 runs such a large region is not sufficiently explored. Nevertheless higher number of conformation sampling, although not assuring better result, would have been too computationally demanding to be pursued.

Assessment of Docking: Re-Docking. In order to check for the reliability of the docking protocol, we performed docking studies on

24 NS5B polymerase complexes. The docking results were evaluated through a comparison of the predicted docked positions of the ligand and the experimental ones. As a measure of docking reliability, the rmsd between the positions of heavy atoms of the ligand in the calculated and experimental structures was considered. The choice of the best conformation was based on the assumption that, although for high-throughput screening protocols only the first ranked docked conformation should be considered (that is, the conformer characterized by the lowest estimated free energy of binding),[248] in other cases, the lowest energy conformer of the most populated cluster should also be taken into account.[249] In Table 6.5 the rmsd for the best docked, best cluster and lowest values obtained are reported. Analyses of these results show that in cases where the best cluster conformer does not coincide with the best docked conformer (1NHU, 1NHV, 1YVZ, 2GIR, 2HWH, 2I1R, 2O5D, 1YVF, 2AWZ, 2AX1, and 2GIQ), a better rmsd value is observed, except for 1NHV, increasing the percentage of conformations found below the threshold value of 2 Å from 41.6% (best docked) to 70.8% (best cluster) to 100% (best fitted cluster). While the best docked is the lowest energy docked conformation of the first Autodock generated cluster, best cluster is the lowest energy docked conformation of the most populated cluster and best fitted cluster is the lowest energy docked conformation of the cluster showing the lowest rmsd value. In the ideal case, i.e. the error free docking program, the three conformations would have coincided. The fact that Autodock reproduces the correct conformation (best fitted cluster) by 100%, reflects the intrinsic lack of accuracy (at least for NS5B) of the implemented scoring function in selecting the right pose, and this is at the basis for the use of the above 3D-QSAR models (as external scoring function) during the virtual screening reported below. As proof of concept we compared

---

[248] Ren, J.; Esnouf, R.; Garman, E.; Somers, D.; Ross, C.; Kirby, I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D.; et al. High resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nat. Struct. Biol.* **1995**, *2*, 293-302.

[249] Ragno, R.; Frasca, S.; Manetti, F.; Brizzi, A.; Massa, S. HIV-reverse transcriptase inhibition: inclusion of ligand-induced fit by cross-docking studies. *J. Med. Chem.* **2005**, *48*, 200-212.

the Autodock predicted pKi values with the experimental pIC$_{50}$ and only a poor correlation was found (data not shown).

Table 6.5. Assessment of the Autodock Program in the Re-Docking Stage. RMSD Values for the First Ranked Pose (Best Docked), the Lowest Energy Docked Conformation of the Most Populated Cluster, (Best Cluster) and the One Closest to the Experimentally Bound Conformation (Best Fitted Cluster).

Binding Site	PDB	Ligand Entry	Best Docked Rmsd	Best Cluster Cluster N°	Best Cluster Rmsd	Best Fitted Cluster Cluster N°	Best Fitted Cluster Rmsd
Thumb	1NHU	**1**	3.17	3	2.17	3	0.89
	1NHV	**2**	4.13	2	4.75	26	1.86
	1OS5	**13**	3.46	1	3.46	9	1.50
	1YVX	**3**	3.81	1	3.81	6	1.58
	1YVZ	**4**	3.74	4	1.92	4	0.78
	2D3U	**6**	0.71	1	0.71	1	0.44
	2D3Z	**7**	0.75	1	0.75	1	0.60
	2D41	**8**	1.43	1	1.43	1	0.58
	2GIR	**5**	5.70	2	1.12	2	0.70
	2HAI	**14**	2.05	1	2.05	1	0.92
	2HWU	**9**	9.86	2	2.13	2	0.79
	2HWI	**10**	0.34	1	0.34	1	0.24
	2I1R	**11**	5.84	2	1.67	2	0.73
	2JC0	**15**	0.85	1	0.85	1	0.68
	2O5D	**12**	5.74	3	2.78	7	1.28
	Average Rmsd		3.44		2.00		0.90
Palm	1YVF	**17**	3.38	4	1.12	4	0.93
	1Z4U	**16**	0.89	1	0.89	1	0.57
	2AWZ	**18**	3.53	3	1.72	3	1.11
	2AX0	**19**	0.84	1	0.84	1	0.52
	2AX1	**20**	3.24	2	0.99	2	0.61
	2FVC	**21**	1.04	1	1.04	1	0.74
	2GC8	**24**	2.03	1	2.03	2	1.82
	2GIQ	**22**	3.49	5	1.91	5	1.81
	2JC0	**15**	0.65	1	0.65	1	0.42
	2JC1	**23**	0.74	1	0.74	1	0.47
	Average Rmsd		1.98		1.19		0.90

Assessment of Docking: Re-Docking Modeled. In addition, the reliability of the docking protocol was evaluated by a re-docking modeled simulation. Our aim was to check more realistically

Autodock's ability to reproduce binding mode conformations of molecules with no experimental data. In this stage therefore, the same modeled conformations used in the Surflex assessment were used to test the Autodock ability. Similarly as reported in the cross-surflex assessment section the docking assessment here was evaluated using as reference conformation the re-surflexed one.

The modeled ligands were submitted to a re-docking simulation using the same docking settings as for the abovementioned re-docking protocol. In Table S5 of the Supporting Information the best docked, best cluster and best fitted cluster rmsd values found are reported. Comparing the values of Table S5 (re-docking modeled) with Table 6.5 (re-docking), at a first glance they do not appear to confirm the ability of Autodock to reproduce the correct binding conformation. However, a visual inspection of the re-docked modeled conformation revealed that using the re-surflexed conformation could lead to a mis-interpretation of the Autodock potentials. In fact upon comparing the three conformations (experimental, re-surflexed and re-docked modeled), a better agreement between the experimental and the re-docked modeled conformations versus the re-surflexed and re-docked modeled one (the tabled rmsd) was noted. As an example of these results, Figure 6.7 depicts the best docked conformations, the reference conformations (Surflex-Sim aligned) and those extracted from the minimized complexes (experimental), for two arbitrarily chosen representatives of thumb (**6**) and palm (**16**) binding NS5B NNIs.
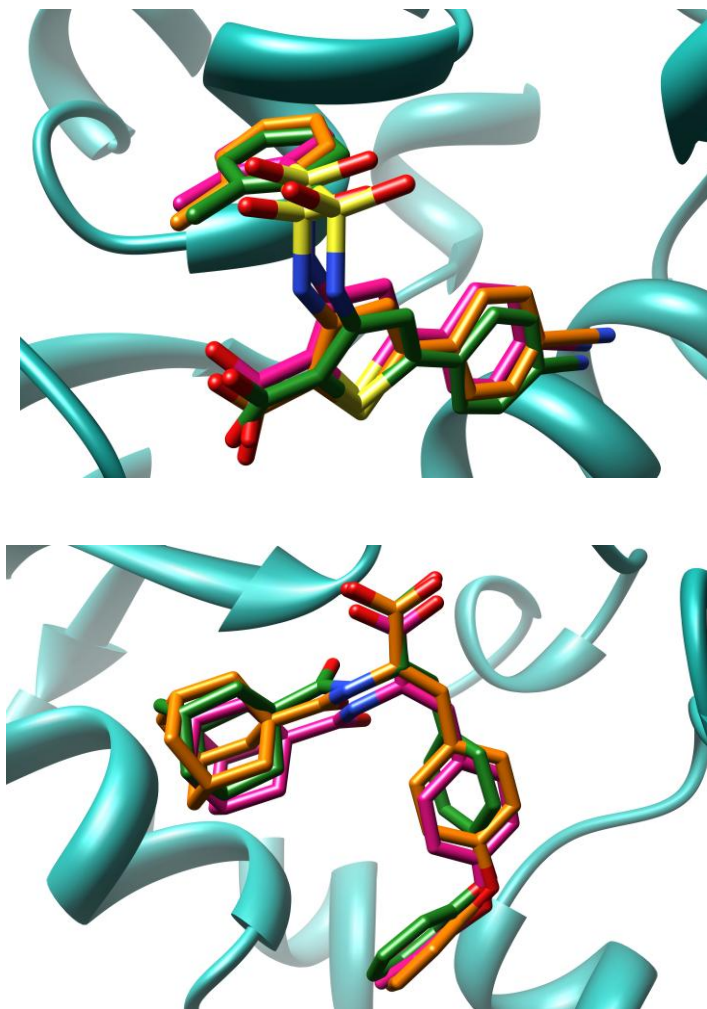
Figure 6.7. Examples of Re-Docking Modeled. Superimposition of the Surflex-aligned conformation (carbon atoms in green) and the re-docked conformation (carbon atoms in magenta) to the experimental conformation (in orange) of one thumb NNI (**6**, on the top) and one palm NNI (**16**, on the bottom) within the NS5B (cyan colored ribbons). Atom bonds are in stick fashion. Hydrogen atoms are omitted for the sake of clarity.

Assessment of Docking: Cross Docking. Along with re-docking experiments (self docking), cross-docking simulations (docking each ligand into the binding pocket of a series of different ligand – receptor complexes) was also performed. The cross-docking method was based on docking each experimental conformation in all the available receptors, while excluding the native one (cross-docking). Then, all output conformations were merged in a single file and clustered. The aim of such a cross-docking procedure was to evaluate the docking program's ability in proposing correct binding conformations, even if no structural data were available for a given modeled and yet untested ligand. In this docking assessment step, similar to the re-docking section, the inclusion of known experimental binding conformation is taken into account.

Additional parameters included evaluation of conformational change during receptor–ligand docking simulations, as well as the consideration of the concept of ligand-induced fit.[249] Autodock 4 was used for all cross-docking calculations employing the same grid boxes parameters (thumb and palm binding sites) and Autogrid maps previously used for re-docking calculations. The numbers of runs of each docking experiment was set so that the number of generated conformations was comparable to that in the re-docking experiments.

In Table 6.6, the rmsd values for the best docked, best cluster and best fitted cluster conformations are reported. As already outlined in the previous paragraph, the best cluster selection versus the best docked criteria gave the best results. Indeed, from analyzing Table 6.6, one can clearly see that the number of correctly docked conformations for the best cluster selection is higher than those obtained by the best docked criteria.

Table 6.6. RMSD Values Obtained by Cross-Docking Simulations

Binding Site	PDB	Ligand Entry	Best Docked	Best Cluster		Best Fitted Cluster	
			Rmsd	Cluster N°	Rmsd	Cluster N°	Rmsd
Thumb	1NHU	**1**	2.25	1	2.25	3	1.87
	1NHV	**2**	4.37	14	3.68	19	2.54
	1OS5	**13**	7.59	5	3.61	32	1.81
	1YVX	**3**	3.51	2	1.77	10	1.58
	1YVZ	**4**	1.73	1	1.73	1	1.67
	2D3U	**6**	5.79	2	5.06	14	1.79
	2D3Z	**7**	5.36	2	3.75	8	1.22
	2D41	**8**	6.49	2	1.51	2	1.51
	2GIR	**5**	1.57	1	1.57	1	1.57
	2HAI	**14**	1.45	1	1.45	1	1.45
	2HWH	**9**	3.60	7	2.9	7	1.71
	2HWI	**10**	3.61	3	2.20	13	1.40
	2I1R	**11**	2.27	2	1.77	3	1.61
	2JC0	**15**	4.26	11	9.7	3	3.38
	2O5D	**12**	3.48	1	3.48	18	1.73
	Average Rmsd		3.82		3.1		1.79
Palm	1YVF	**17**	2.56	2	3.43	9	2.52
	1Z4U	**16**	2.26	2	3.87	12	1.78
	2AWZ	**18**	1.64	1	1,.64	1	1.64
	2AX0	**19**	1.85	1	1.85	3	1.58
	2AX1	**20**	1.68	1	1.68	1	1.68
	2FVC	**21**	1.88	1	1.88	3	1.06
	2GC8	**24**	11.13	2	2.08	6	1.84
	2GIQ	**22**	1.84	1	1.84	11	1.64
	2JC0	**15**	1.21	1	1.21	1	1.21
	2JC1	**23**	0.97	1	0.97	1	0.97
	Average Rmsd		2.7		2.09		1.59

Assessment of Docking: Cross-Docking Modeled. Cross-docking experiments of modeled training sets ligands were also carried out. Grid box parameters and all docking settings were the same as for the cross-docking simulations. In Table S6 of the Supporting Information, the rmsd for the best docked, best cluster and lowest rmsd values found are reported. As outlined above, we have evaluated the ability of the docking program ability in proposing correct binding conformations even when no structural

data are available for a given modeled and yet untested ligand. Here the docking program is tested as in a real case situation where a randomly chosen ligand conformation is being docked into a receptor binding pocked whose conformation do not bear any information on the actual ligand inducing fit.

Table S6 again confirms that the best cluster conformation have to be considered instead of the best docked. Here again similarly, though it is less evident compared to the re-docking modeled assessment, the rmsd values seems to mis-lead in the abilities of Autodock to correctly reproduce acceptable docked conformations. Visual inspection (data not shown) revealed that there is a better agreement between the experimental and the cross-docked modeled conformations, than those observed between the re-surflexed and cross-docked modeled conformation (rmsd values in Tables S5 and S6).

Similarly as stated above for the cross-surflex experiments, in this last SB alignment assessment, the ability to accurately align randomly built conformations, suggested that Autodock can be used as an useful tool for SB alignment of NS5B inhibitors (see external test set below), and to some extent to align molecule in a virtual screening protocols (see below). A schematic description of the workflow of the alignment process is shown in Figure 6.8.
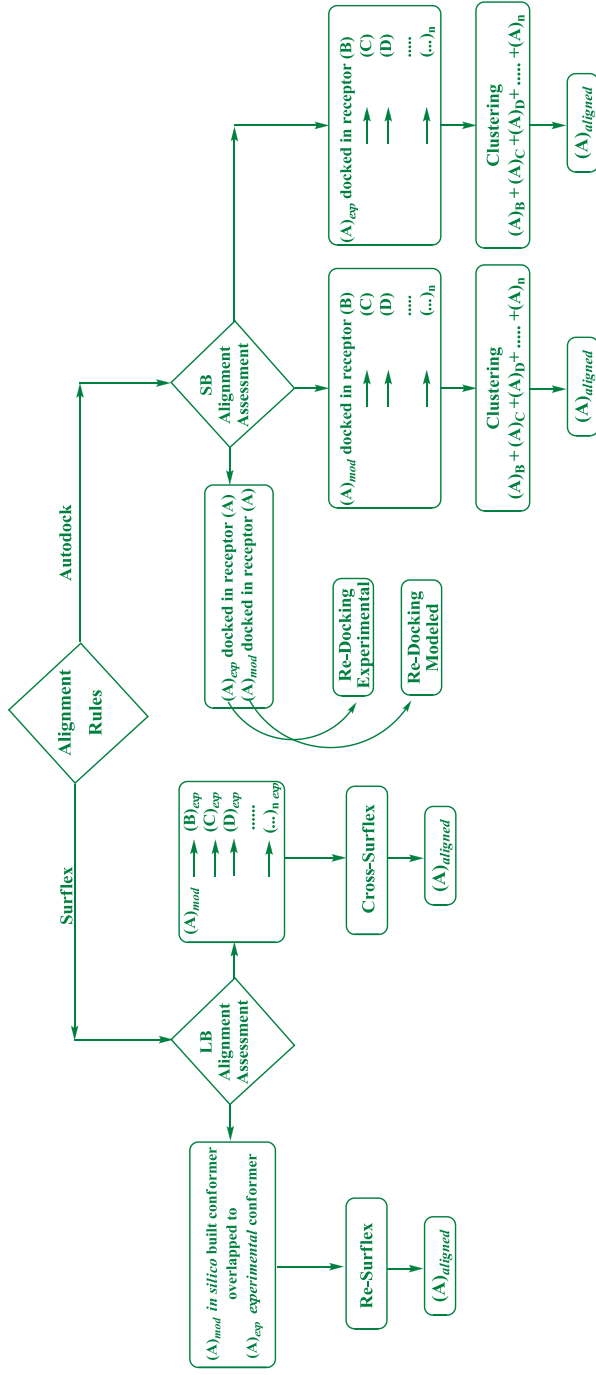
Figure 6.8. Description of the workflow of the alignment protocol. Key steps for the assessment of LB and SB alignments processes.

6.3. External Validation of the 3D-QSAR Models

To evaluate the predictive ability of the 3D-QSAR models, two series of test sets were used. Compounds of the first test set include 81 NS5B NNIs selected from literature, that bind in the thumb subdomain of the receptor. Molecular structures of thumb test set are represented in Table S8 – Table S19 of the Supporting Information. Compounds of the second test set include 223 NS5B ligands reported in literature as palm subdomain allosteric inhibitors. Molecular structures of palm test set are represented in Table S20 – Table S36 of the Supporting Information. Test set compounds were selected so that they span the whole descriptor (chemical) space occupied by the training set and cover as much as possible the same range of bioactivities.

All 304 test set molecules were generated by means of molecular mechanics of Chemaxon Marvin software (http://www.chemaxon.com/) and all molecular graphics images were produced using UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco on a 3 Ghz AMD CPU equipped IBM compatible workstation with the Debian 5.0 version of the Linux operating system. Then, test set compounds were submitted to the molecular alignment using the LB (cross-surflex) and SB (cross-docking) protocols as reported above. In the alignment of the external test sets all the experimental ligands were used as target list (for each receptor site) in Surflex and all the receptor structures were used in the cross-docking experiments.

Each test set compounds was then aligned twice (cross-surflex conformation and best cluster conformation from cross-docking) and the respective MIFs calculated.

Application of the two SB 3-D QSAR models to the external test sets (C1= GRID probe) yielded low standard deviation error of predictions values, shown in Table S1–Table S2 of the Supporting Information for all GRID probes.

Analysis of individual error of predictions are graphically reported in Figure 6.9. Here, it is clearly visible how the 3D-QSAR models, coupled to LB and SB alignment methods, can be used as external scoring function. In particular, the use of quite large external test sets is a sort of a virtual screening simulation (although all molecules are known to bind NS5B). In particolar, from Figure 6.9, it is possible to observe a good agreement between the two alignment methods, and setting to 1 an arbitrarily threshold value of acceptable prediction error, the percentage of compounds well predicted ranged from 54% to 63% and rose to a maximum of 92% if the cutoff value would set to 1.5.

The interesting results of the above mixed protocol prompted us to undertake a virtual screening approach, which is reported in the next section. It has to be pointed up that either training set or test set compounds were biologically tested in similar condition assays.
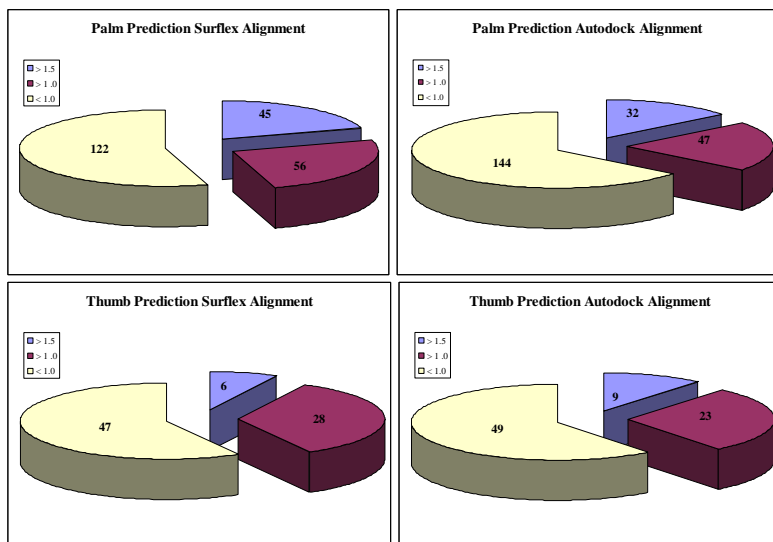
Figure 6.9. Diagrams of number of compounds with an error of prediction greater than 1.5 (violet), 1.0 (dark red) and lower than 1.0 (pale yellow).

6.4. Virtual Screening

Structure-based virtual screening has become an integral part of the drug design process and is now the most widely-used approach to leverage structure for ligand discovery. In receptor-based virtual screening, the main purpose is to find a ligand that has a high binding affinity to the target receptor whose 3D structure is known. Ligand similarity searches also provide a very powerful method of quickly screening large databases of ligands to identify possible hits with activity against the target of interest. There are many tools available for performing these computational analyses and they can be categorized as being either ligand-based or receptor-based. For ligand-based methods, the strategy is to perform similarity search if one or more active molecules are known. This can be done by a variety of methods, QSAR and 3D-QSAR models as well as pharmacophore matching.

Receptor-based computational methods can be employed when the 3D structure of the target protein is known. These involve docking algorithms to put ligands into the binding site of the target, and thus producing a predicted binding mode for each database compound. The application of scoring functions which in turn produce a ranking list of compounds is necessary to identify potential hits.

The VS protocol reported in this study is based on the application of two receptor-based 3D-QSAR models, on external test sets containing 81 molecules for the thumb model and 223 molecules for the palm model (see Figure 6.10 for a detailed description of the VS workflow). This procedure proved them to be useful scoring functions for the virtual screening approach, using either a ligand-based (Surflex-Sim) or a structure-based (Autodock) molecular alignment. Consequently, the NCI Diversity Set (DS) (http://dtp.nci.nih.gov), a database consisting of over 1990 compounds, was virtually screened. For each statistical model, the first 20 most active predicted molecules were selected for biological assays against recombinant NS5BCΔ21 (genotype 1b).
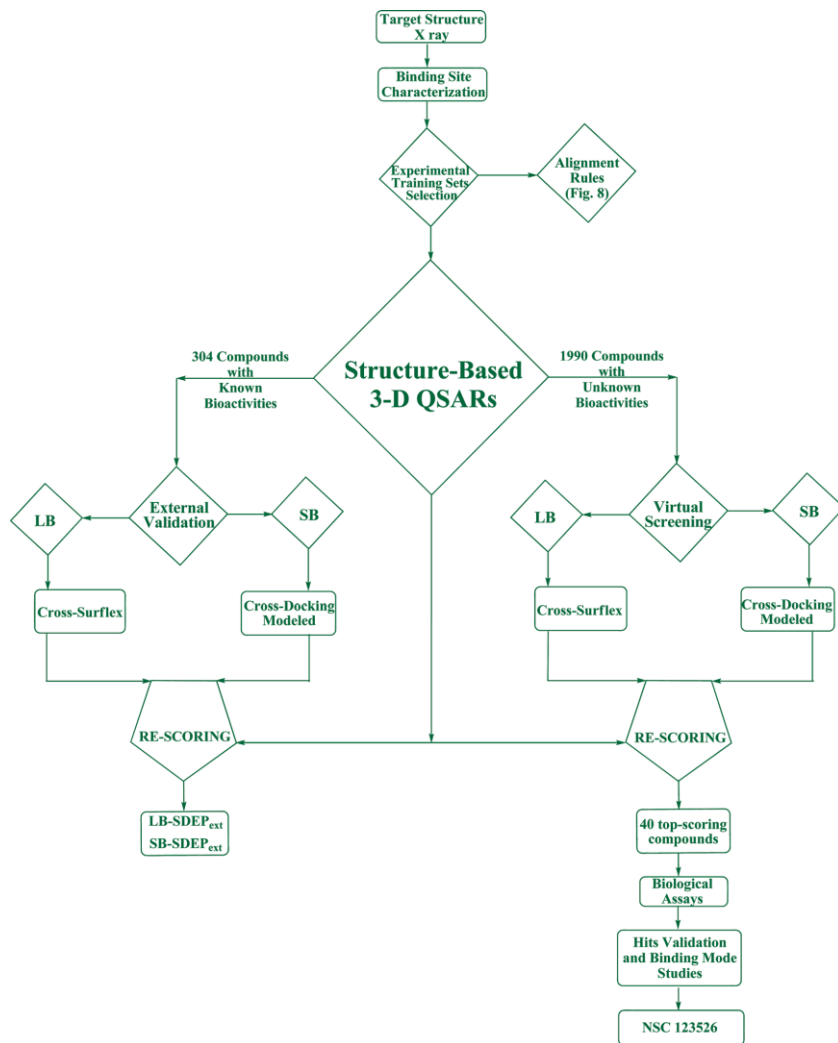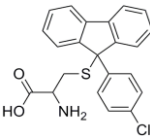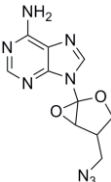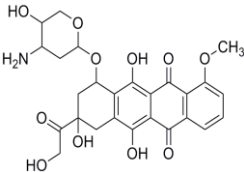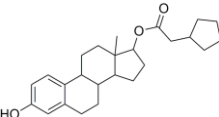
Figure 6.10. Workflow of 3D-QSAR studies and alignment rules integration in the virtual screening process used in the present work.

Among the selected compounds, preliminary data yielded four derivatives (Table 6.7) with IC_{50} values between 45-75 μM. The selected compounds have been tested for their activity using

similar biological conditions as for training set and external test set compounds (see biological data). It could be argued that the NCI DS may have limitations to screen for compounds active against NS5B. However, the NCI database is one of the most widely used as it exhibits molecular diversity. Moreover the implicit limitation of such a choice further strengths the applied approach considering that 10% of the tested molecules were indeed found active in enzyme-based assay.

Table 6.7. Molecular Structure and Antiviral Activity[a] of the Compounds Selected by VS Protocol

Thumb Domain		Palm Domain	

NSC 123526	NSC 125626	NSC 169534	NSC 3354
IC$_{50}$ = 46.0 μM	IC$_{50}$ = 73.3 μM	IC$_{50}$ = 64.5 μM	IC$_{50}$ = 54.3 μM

[a] The data represents an average of at least two independent experiments.

6.4.1.   Biological data: NS5B RdRp Assay

Recombinant NS5BCΔ21 was purified from the plasmid pThNS5BCΔ21 expressed in *Escherichia coli* DH5α by Ni-NTA affinity chromatography, and employed for evaluating the inhibitory activity of the compounds on poly rA/U$_{12}$ template-primer (TP) as previously reported.[250,251] NS5B RdRp activity was determined in the absence or presence of the compounds at 30°C for 60 min in a

[250] Kaushik-Basu, N.; Bopda-Waffo, A.; Talele, T. T.; Basu, A.; Costa, P. R.; da Silva, A. J.; Sarafianos, S. G.; Noel, F. Identification and characterization of coumestans as novel HCV NS5B polymerase inhibitors. *Nucleic Acids Res.* **2008**, *36*, 1482-1496.

[251] Kaushik-Basu, N.; Bopda-Waffo, A.; Talele, T. T.; Basu, A.; Chen, Y.; Kucukguzel, S. G. 4-Thiazolidinones: a novel class of hepatitis C virus NS5B polymerase inhibitors. *Front. Biosci.* **2008**, *13*, 3857-3868.

reaction buffer (25 μL) containing 20 mM Tris–HCl [pH 7.0], 100 mM NaCl, 100 mM sodium glutamate, 0.5 mM DTT, 0.01% BSA, 0.01% Tween-20, 5% glycerol, 20 U/mL of RNase Out, 0.5 μM poly rA/U$_{12}$ (pre-annealed), 25 μM UTP, 2–5 μCi [α-^{32}P]UTP, 0.5 mM MnCl$_2$ and 300 ng of NS5BCΔ21. Reactions were terminated by the addition of ice-cold 10% (v/v) trichloroacetic acid (TCA) containing 0.5 mM pyrophosphate. The quenched reaction mixtures were spotted on GF-B filters, washed with 5% (v/v) TCA-0.5 mM pyrophosphate to remove unincorporated rNTPs, rinsed and quantified on a liquid scintillation counter (Packard). Activity of NS5B in the absence of the inhibitor was set at 100% and that in the presence of the inhibitor was calculated relative to this control. The 50% inhibition value (IC$_{50}$) for the inhibitors was evaluated from the dose response curve employing typically 9-12 concentrations of the compound and GraphPad Prism 5.0 soft. N,N-disubstituted phenylalanine derivative 14 (as numbered in the original paper),[207] a documented NS5B inhibitor was included as positive control in each set of experiments.

6.4.2. Binding Mode Analysis

Analysis of the binding mode of selected compounds (Table 6.7) within the NS5B RdRp polymerase thumb domain was performed with the aim of elaborating a putative way by which they inhibit the enzyme, and thus, getting insights on the peculiar chemical features that influenced their activity. Among the four active compounds, only one draws (**NSC 123526**) particular attention from a medicinal chemistry point of view. In fact, compound **NSC 125626** is an oxirane (thus likely an irreversible inhibitor) and its structure is similar to a nucleoside, thus indicating it as a nucleoside inhibitor; compounds **NSC 169534** and **NSC 3354** are structurally too rigid and structurally resemble a tetracycline and a colestane related compound, respectively. As a matter of fact, the resulting four complexes were visually inspected, and only **NSC 123526** endowed with the lowest inhibitory activity (IC$_{50}$ = 46.0 μM), was found in good agreement with the orientation

of the most active thumb binding compound **6**. Moreover, a search for compound **NSC 123526** in literature highlighted its effective activity against other polymerases diverse from NS5B RdRp.[252] Figure 6.11 shows the predicted binding mode of the minimum energy conformation of the hit compound **NSC 123526** and the experimental conformation of the thumb inhibitor **6**. In detail, the para-chlorine-phenyl ring of **NSC 123526** is located in the region occupied by the ciano-phenyl-thiophene moiety of inhibitor **6** and characterized by van der Waals interactions with Leu419, Met423, Ile482, Ala486, Leu489 and Leu497. As reported,[213] the ciano-phenyl-thiophene moiety (**6**) is bound to a hydrophobic surface of the thumb domain. In a similar fashion to the 4-methyl-phenyl group of inhibitor **6**, the tricyclic ring system of compound **NSC 123526** points towards the interior of the thumb domain, filling the primary binding cavity by making extensive van der Waals interactions with Leu419, Met423, Leu474, His475, and Trp528. In addition, one of the carboxylic oxygen atoms of NSC 123526 may be involved in a hydrogen bond with the guanidinium group of Arg501. This is in full agreement with experimental data that show one of the sulphonyl oxygen atoms of compound **6** bonded to the same group of Arg501.

---

[252] Silverman, J. E.; Ciustea, M.; Shudofsky, A. M.; Bender, F.; Shoemaker, R. H.; Ricciardi, R. P. Identification of polymerase and processivity inhibitors of vaccinia DNA synthesis using a stepwise screening approach. *Antiviral Res.* **2008**, *80*, 114-123.
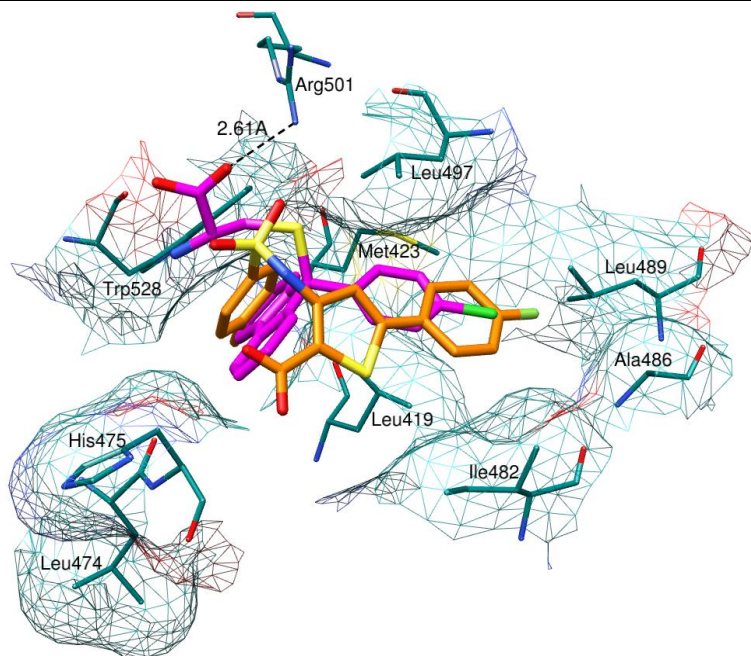
Figure 6.11. Superimposition of inhibitor **6** (orange) and hit compound **NSC 123526** (magenta) structures, showing a comparable mode of binding within the NS5B thumb domain. The residues in the binding site are shown by stick representation with carbon, nitrogen, oxygen, and sulfur atoms in cyan, blue, red, and yellow, respectively. For the sake of clarity, hydrogen bond interactions are represented by black dotted lines, and only a few residues of the binding site were displayed. Hydrogen atoms are omitted.

6.5. Conclusions

Two structure-based 3D QSAR models coupled with both ligand-based and structure-based alignment methods were developed for NS5B non-nucleoside inhibitors. The overall protocol was extensively assessed for both alignment and predictive ability. As final test a virtual screening application using the NCI Diversity Set 1990 compounds led to the discovery of a new active compound (**NSC 123526**) with an unprecedented scaffold among NS5B known inhibitors. Further virtual screening are in due

progress on wider and more appropriate containing structures database (i.e. Maybridge), furthermore a focused virtual screening will be set up on a library of compounds structurally related with **NSC 123526**.

6.6. Supporting Information Available

Statistical results for the thumb and the palm training sets analyzed by different GRID probes, Rmsd values obtained by re-docking simulations of modeled compounds, Rmsd values obtained by cross-docking simulations of modeled compounds, results from the assessment of Surflex-Sim. Molecular structures of 223 NNIs used as test set for palm TS Sim LB alignment protocol and molecular structures of 223 NNIs used as test set for palm TS.

Table S1. Statistical Results Obtained from GOLPE PLS Analysis[a] for the Thumb and the Palm Training Sets

Training Set	P	Ligand-Based Alignment[b]				
		r^2	q^2	$SDEP_{LSO-5}$	PC	$LB\text{-}SDEP_{ext}$[b]
Thumb	C1=	0.99	0.69	0.31	3	0.65
	C3	0.99	0.59	0.36	3	0.68
	DRY	0.98	0.36	0.45	3	0.69
	OH	0.99	0.60	0.36	3	0.66
	O::	0.99	0.45	0.43	3	0.60
	N1=	0.99	0.51	0.40	3	0.72
	N3+	0.99	0.50	0.40	3	0.72
	N1	0.99	0.64	0.34	3	0.67
Palm	C1=	0.99	0.55	0.66	3	1.03
	C3	0.99	0.51	0.68	3	1.06
	DRY	0.99	0.03	0.96	3	0.94
	OH	0.99	0.49	0.68	3	1.09
	O::	0.99	0.57	0.64	3	1.17
	N1=	0.99	0.59	0.62	3	1.39
	N3+	0.98	0.56	0.64	3	1.38

[a] P, GRID probe; r^2, conventional square correlation coefficient; q^2, cross-validation correlation coefficient; $SDEP_{LSO-5}$, cross-validated standard error of prediction using leave-some-out cross-validation using five groups; PC, optimal number of principal components; $SDEP_{ext}$, standard error of prediction for the external test set.
[b] LB-SDEP: SDEP values obtained using external test set molecules used for preliminary statistical analysis that were superimposed using a ligand-based procedure via Surflex-Sim program.

Table S2. Statistical Results Obtained from GOLPE PLS Analysis[a] for the Thumb and the Palm Training Sets

Training Set	P	r^2	q^2	SDEP	PC	SB-SDEP$_{ext}$[b]
			Structure-Based Alignment[b]			
Thumb	C1=	0.99	0.69	0.31	3	0.59
	C3	0.99	0.59	0.36	3	0.60
	DRY	0.98	0.36	0.45	3	0.77
	OH	0.99	0.60	0.36	3	0.58
	O::	0.99	0.45	0.43	3	0.59
	N1=	0.99	0.51	0.40	3	0.59
	N3+	0.99	0.50	0.40	3	0.59
	N1	0.99	0.64	0.34	3	0.58
Palm	C1=	0.99	0.55	0.66	3	1.08
	C3	0.99	0.51	0.68	3	1.08
	DRY	0.99	0.03	0.96	3	1.21
	OH	0.99	0.49	0.68	3	1.09
	O::	0.99	0.57	0.64	3	1.11
	N1=	0.99	0.59	0.62	3	1.14
	N3+	0.98	0.56	0.64	3	1.14

[a] P, GRID probe; r^2, conventional square correlation coefficient; q^2, cross-validation correlation coefficient; SDEP$_{LSO-5}$, cross-validated standard error of prediction using leave-some-out cross-validation using five groups; PC, optimal number of principal components; SDEP$_{ext}$, standard error of prediction for the external test set.
[b] SB-SDEP: SDEP values obtained using external test set molecules used for preliminary statistical analysis that were superimposed using a structure-based procedure via Autodock4 program.

Table S3. List of entries and biological activities of compounds of the preliminary thumb test set used in the first steps of the work in order to determine the GRID probe for model interpretation and the optimal number of principal components

Entry[a]	Ref	exp IC$_{50}$	pred.pIC$_{50}$[b]	pred.pIC$_{50}$[c]
4	1	5.7	5.51	6.09
7	1	5.92	5.76	5.39
8	2	6.12	5.6	6.41
9	2	6.0	5.48	6.21
11	2	6.41	5.84	6.27
12	1	5.85	5.78	5.24
13	2	4.7	5.31	6.12
17	1	4.62	5.45	5.69
23	3	5.54	5.71	6.4
24	3	5.77	5.97	6.29
30	3	6	5.77	5.88
32	3	6.1	5.84	5.98
35	3	6.25	5.79	5.96
39	3	7.42	5.44	6.04
9	4	5.1	5.49	5.64
11	4	5.49	5.89	5.67
14	4	6.15	5.69	5.62
15	4	6.1	5.8	5.56
17	4	5.1	5.71	5.64
9	5	5.18	5.57	5.31
11	5	4.85	5.45	5.54

[a] Compounds are numbered as in their original papers. [b] *SB alignment* [c] *LB alignment*

Table S4. List of entries and biological activities of compounds of the preliminary palm test set used in the first steps of the work in order to determine the GRID probe for model interpretation and the optimal number of principal components

Entry[a]	Ref	exp.pIC$_{50}$	pred.pIC$_{50}$	pred.pIC$_{50}$
28	6	7.3	5.49	6.86
37	6	6.09	5.9	7.1
38	6	5.41	5.79	6.84
44	7	5.06	5.84	7.11
46	6	5.55	6.29	6.84
48	7	4.7	5.62	6.79
51	6	5.15	6.25	6.96
57	6	6.1	5.64	6.72
58	7	7.4	5.82	7.06
61	7	5.89	6.34	7.09
11	8	8.25	6.7	7.65
25	9	7.1	5.61	5.49
2e	10	4.8	6.19	5.98
39	8	7.28	6.87	7.49
5aa	10	6.22	6.61	6.19
5e	10	5.3	6.98	6.04
5l	10	6.05	6.24	6.18
5u	10	6.7	6.58	6.03
5v	10	5.1	6.58	5.8
68	8	7.57	6.85	7.5
73	8	7.48	6.98	7.54
90	8	8.4	6.71	7.75
9b	11	7.85	6.53	7.09

[a] Compounds are numbered as in their original papers. [b] SB alignment [c] LB alignment

References

[1]Chan, L.; Pereira, O.; Reddy, T. J.; Das, S. K.; Poisson, C.; Courchesne, M.; Proulx, M.; Siddiqui, A.; Yannopoulos, C. G.; Nguyen-Ba, N.; Roy, C.; Nasturica, D.; Moinet, C.; Bethell, R.; Hamel, M.; L'Heureux, L.; David, M.; Nicolas, O.; Courtemanche-Asselin, P.; Brunette, S.; Bilimoria, D.; Bedard, J. Discovery of thiophene-2-carboxylic acids as potent inhibitors of HCV NS5B polymerase and HCV subgenomic RNA replication. Part 2: tertiary amides. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 797-800.

[2]Chan, L.; Das, S. K.; Reddy, T. J.; Poisson, C.; Proulx, M.; Pereira, O.; Courchesne, M.; Roy, C.; Wang, W.; Siddiqui, A.; Yannopoulos, C. G.; Nguyen-Ba, N.; Labrecque, D.; Bethell, R.; Hamel, M.; Courtemanche-Asselin, P.; L'Heureux, L.; David, M.; Nicolas, O.; Brunette, S.; Bilimoria, D.; Bedard, J. Discovery of thiophene-2-carboxylic acids as potent inhibitors of HCV NS5B polymerase and HCV subgenomic RNA replication. Part 1: Sulfonamides. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 793-796.

[3]Li, H.; Tatlock, J.; Linton, A.; Gonzalez, J.; Borchardt, A.; Dragovich, P.; Jewell, T.; Prins, T.; Zhou, R.; Blazel, J.; Parge, H.; Love, R.; Hickey, M.; Doan, C.; Shi, S.; Duggal, R.; Lewis, C.; Fuhrman, S. Identification and structure-based optimization of novel dihydropyrones as potent HCV RNA polymerase inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4834-4838.

[4]Chan, L.; Reddy, T. J.; Proulx, M.; Das, S. K.; Pereira, O.; Wang, W.; Siddiqui, A.; Yannopoulos, C. G.; Poisson, C.; Turcotte, N.; Drouin, A.; Alaoui-Ismaili, M. H.; Bethell, R.; Hamel, M.; L'Heureux, L.; Bilimoria, D.; Nguyen-Ba, N. Identification of N,N-disubstituted phenylalanines as a novel class of inhibitors of hepatitis C NS5B polymerase. *J. Med. Chem.* **2003**, *46*, 1283-1285.

[5]Yan, S.; Appleby, T.; Larson, G.; Wu, J. Z.; Hamatake, R. K.; Hong, Z.; Yao, N. Thiazolone-acylsulfonamides as novel HCV NS5B polymerase allosteric inhibitors: convergence of structure-based drug design and X-ray crystallographic study. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1991-1995.

[6]Pfefferkorn, J. A.; Nugent, R.; Gross, R. J.; Greene, M.; Mitchell, M. A.; Reding, M. T.; Funk, L. A.; Anderson, R.; Wells, P. A.; Shelly, J. A.; Anstadt, R.; Finzel, B. C.; Harris, M. S.; Kilkuskie, R. E.; Kopta, L. A.; Schwende, F. J. Inhibitors of HCV NS5B polymerase. Part 2: Evaluation of the northern region of (2Z)-2-benzoylamino-3-(4-phenoxy-phenyl)-acrylic acid. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2812-2818.

[7]Pfefferkorn, J. A.; Greene, M. L.; Nugent, R. A.; Gross, R. J.; Mitchell, M. A.; Finzel, B. C.; Harris, M. S.; Wells, P. A.; Shelly, J. A.; Anstadt, R. A.; Kilkuskie, R. E.; Kopta, L. A.; Schwende, F. J. Inhibitors of HCV NS5B polymerase. Part 1: Evaluation of the southern region of (2Z)-2-(benzoylamino)-3-(5-phenyl-2-furyl)acrylic acid. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2481-2486.

[8] Tedesco, R.; Shaw, A. N.; Bambal, R.; Chai, D.; Concha, N. O.; Darcy, M. G.; Dhanak, D.; Fitch, D. M.; Gates, A.; Gerhardt, W. G.; Halegoua, D. L.; Han, C.; Hofmann, G. A.; Johnston, V. K.; Kaura, A. C.; Liu, N.; Keenan, R. M.; Lin-Goerke, J.; Sarisky, R. T.; Wiggall, K. J.; Zimmerman, M. N.; Duffy, K. J. 3-(1,1-dioxo-2H-(1,2,4)-benzothiadiazin-3-yl)-4-hydroxy-2(1H)-quinolinones , potent inhibitors of hepatitis C virus RNA-dependent RNA polymerase. *J. Med. Chem.* **2006**, *49*, 971-983.

[9] Gopalsamy, A.; Chopra, R.; Lim, K.; Ciszewski, G.; Shi, M.; Curran, K. J.; Sukits, S. F.; Svenson, K.; Bard, J.; Ellingboe, J. W.; Agarwal, A.; Krishnamurthy, G.; Howe, A. Y.; Orlowski, M.; Feld, B.; O'Connell, J.; Mansour, T. S. Discovery of proline sulfonamides as potent and selective hepatitis C virus NS5b polymerase inhibitors. Evidence for a new NS5b polymerase binding site. *J. Med. Chem* **2006**, *49*, 3052-3055.

[10] Powers, J. P.; Piper, D. E.; Li, Y.; Mayorga, V.; Anzola, J.; Chen, J. M.; Jaen, J. C.; Lee, G.; Liu, J.; Peterson, M. G.; Tonn, G. R.; Ye, Q.; Walker, N. P.; Wang, Z. SAR and mode of action of novel non-nucleoside inhibitors of hepatitis C NS5b RNA polymerase. *J. Med. Chem.* **2006**, *49*, 1034-1046.

[11] Slater, M. J.; Amphlett, E. M.; Andrews, D. M.; Bravi, G.; Burton, G.; Cheasty, A. G.; Corfield, J. A.; Ellis, M. R.; Fenwick, R. H.; Fernandes, S.; Guidetti, R.; Haigh, D.; Hartley, C. D.; Howes, P. D.; Jackson, D. L.; Jarvest, R. L.; Lovegrove, V. L.; Medhurst, K. J.; Parry, N. R.; Price, H.; Shah, P.; Singh, O. M.; Stocker, R.; Thommes, P.; Wilkinson, C.; Wonacott, A. Optimization of novel acyl pyrrolidine inhibitors of hepatitis C virus RNA-dependent RNA polymerase leading to a development candidate. *J. Med. Chem.* **2007**, *50*, 897-900.

Table S5. RMSD Values Obtained by Re-Docking Simulations of Modeled Compounds

Binding Site	PDB	Ligand Entry	Best Docked			Best Cluster			Best Fitted Cluster	
			Rmsd	Cluster N°	Rmsd	Cluster N°	Rmsd		Cluster N°	Rmsd
	1NHU	**1**	3.13			1	3.13		1	3.13
	1NHV	**2**	4.08			1	4.08		7	3.44
	1OS5	**13**	4.41			2	2.98		14	1.97
	1YVX	**3**	3.93			1	3.93		4	3.66
	1YVZ	**4**	4.03			1	4.03		20	1.98
	2D3U	**6**	1.88			4	3.36		1	1.85
	2D3Z	**7**	3.70			7	5.20		1	3.70
Thumb	2D41	**8**	2.80			4	11.01		1	2.80
	2GIR	**5**	5.53			5	2.83		5	2.83
	2HAI	**14**	1.94			1	1.94		1	1.22
	2HWH	**9**	10.03			3	3.27		3	3.16
	2HWI	**10**	1.76			2	10.66		1	1.76
	2I1R	**11**	11.05			2	3.29		2	3.14
	2JC0	**15**	3.83			3	11.02		1	1.57
	2O5D	**12**	6.24			2	6.54		5	4.72
	average		4.56				5.15			2.73
Palm	1YVF	**17**	3.60			1	3.60		1	3.32
	1Z4U	**16**	3.47			1	3.47		12	2.09

NS5B RESULTS ANALYSIS

Binding Site	PDB	Ligand Entry	Best Docked	Best Cluster		Best Fitted Cluster	
			Rmsd	Cluster N°	Rmsd	Cluster N°	Rmsd
	2AWZ	**18**	5.18	3	2.68	3	1.47
	2AX0	**19**	4.62	2	4.48	4	3.30
	2AX1	**20**	5.40	6	5.85	9	3.84
	2FVC	**21**	3.78	15	2.64	15	2.24
Palm	2GC8	**24**	2.26	1	2.26	1	1.22
	2GIQ	**22**	3.69	1	3.69	2	2.48
	2JC0	**15**	1.53	1	1.53	1	1.18
	2JC1	**23**	2.27	1	2.27	1	1.79
	average		3.58		3.25		2.29
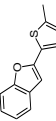
Table S6. RMSD Values Obtained by Cross-Docking Simulations of Modeled Compounds

Binding Site	PDB	Ligand Entry	Best Docked		Best Cluster			Best Fitted Cluster	
			Rmsd		Cluster N°	Rmsd		Cluster N°	Rmsd
	1NHU	**1**	8.82		3	3.43		4	1.90
	1NHV	**2**	3.81		2	3.84		17	2.49
	1OS5	**13**	1.98		9	3.51		1	1.98
	1YVX	**3**	1.72		1	1.72		1	1.72
	1YVZ	**4**	2.84		1	2.84		5	1.81
	2D3U	**6**	6.61		2	2.22		35	1.52
	2D3Z	**7**	2.15		1	2.15		6	0.96
Thumb	2D41	**8**	8.47		11	2.12		14	1.60
	2GIR	**5**	1.29		1	1.29		1	1.29
	2HAI	**14**	1.32		1	1.32		1	1.32
	2HWH	**9**	3.24		3	8.69		5	1.61
	2HWI	**10**	4.59		1	4.59		33	2.19
	2I1R	**11**	2.78		1	2.78		8	1.69
	2JC0	**15**	11.08		6	8.99		5	2.40
	2O5D	**12**	4.78		7	3.37		20	2.34
	average		4.37			3.52			1.79
Palm	1YVF	**17**	2.78		1	2.78		7	1.95
	1Z4U	**16**	2.90		1	2.90		20	2.16

NS5B RESULTS ANALYSIS

Binding Site	PDB	Ligand Entry	Best Docked	Best Cluster		Best Fitted Cluster	
			Rmsd	Cluster N°	Rmsd	Cluster N°	Rmsd
	2AWZ	**18**	1.62	1	1.62	1	1.62
	2AX0	**19**	2.85	1	2.85	16	1.91
	2AX1	**20**	5.43	2	2.85	7	2.04
	2FVC	21	1.76	1	1.76	4	1.63
	2GC8	24	2.87	1	2.87	12	1.58
Palm	2GIQ	22	1.71	1	1.71	1	1.71
	2JC0	15	1.74	1	1.74	1	1.37
	2JC1	23	1.87	1	1.87	2	1.58
average			2.55		2.3		1.76

Table S7. Assessment of Surflex-Sim LB Alignment Protocol. RMSD Values Between the Modeled Conformation of Either Training Sets Compounds and the Reference Compound.

	PDB	Entry	RMSD		PDB	Entry	RMSD
Thumb Inhibitors	1NHU	**1**	0.51	**Palm Inhibitors**	1YVF	**17**	0.74
	1NHV	**2**	0.79		1Z4U	**16**	0.82
	1OS5	**13**	0.58		2AWZ	**18**	0.60
	1YVX	**3**	0.65		2AX0	**19**	0.56
	1YVZ	**4**	0.87		2AX1	**20**	0.52
	2D3U	**6**	0.85		2FVC	**21**	0.70
	2D3Z	**7**	0.83		2GC8	**24**	0.43
	2D41	**8**	0.80		2GIQ	**22**	0.67
	2GIR	**5**	0.65		2JC0	**15**	0.80
	2HAI	**14**	0.51		2JC1	**23**	0.78
	2HWH	**9**	0.68				
	2HWI	**10**	0.69				
	2I1R	**11**	0.66				
	2JC0	**15**	0.36				
	2O5D	**12**	0.79				

NS5B RESULTS ANALYSIS

Table S8. *N,N*-disubstituted Phenylalanine Derivatives.[a]
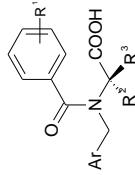
Compd.	Paper no.[a]	R[1]	Ar	R[2]	R[3]	exp. pIC$_{50}$	pred pIC$_{50}$[b]	pred. pIC$_{50}$[c]
1	1	2,4-Cl$_2$	(thiazol-tolyl)	CH$_2$Ph	H	5.24	5.98	5.76
2	2	2,4-Cl$_2$	2-BrPh	CH$_2$Ph	H	5.10	5.96	5.57
3	3	2,4-Cl$_2$	3-BrPh	CH$_2$Ph	H	5.46	5.52	5.68
4	4	2,4-Cl$_2$	4-BrPh	CH$_2$Ph	H	5.04	5.36	5.69
5	5	2,4-Cl$_2$	3-BrPh	H	H	4.28	5.72	5.63
6	6	2,4-Cl$_2$	3-BrPh	H	H	5.07	5.06	5.67
7	7	2,4-Cl$_2$	3-BrPh	CH$_2$Ph	H	4.96	5.72	5.58
8	8	2,4-Cl$_2$	3-BrPh	CH$_2$Ph	H	4.82	5.71	6.10
9	10	2,4-Cl$_2$	3-ClPh	CH$_2$Ph	H	5.85	5.66	5.69
10	12	2,4-Cl$_2$	3-MePh	CH$_2$Ph	H	5.74	6.00	5.63
11	13	2,4-Cl$_2$	3-CF$_3$Ph	CH$_2$Ph	H	5.77	5.95	5.72
12	16	2,4-Cl$_2$	2-CNPh	CH$_2$Ph	H	5.22	5.79	5.64
13	18	2,4-Cl$_2$	4-pyridyl N-oxide	CH$_2$Ph	H	5.57	6.11	5.65
14	19	2,4-Cl$_2$	3-pyridyl	CH$_2$Ph	H	4.96	5.70	5.57

NS5B RESULTS ANALYSIS

Compd.	Paper no.[a]	R^1	Ar	R^2	R^3	exp. pIC_{50}	pred pIC_{50}[b]	pred. pIC_{50}[c]
15	**20**	2,4-Cl_2		CH_2Ph	H	5.51	5.50	5.70
16	**21**	2,4-Cl_2		CH_2Ph	H	5.07	5.71	5.79

[a] Journal of Medicinal Chemistry 2003, 46, 1283-1285 [b] SB alignment [c] LB alignment

Table S9. Dihydropyrones Derivatives.[a]
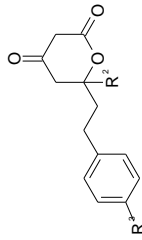
Compd.	Paper no.[a]	R^2	R^3	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
17	**18**	Cyclobutyl	OH	4.03	6.00	5.70
18	**19**	Cyclopentyl	OH	5.09	5.99	6.30
19	**20**	Cyclohexyl	OH	4.28	6.27	5.86

[a] *Bioorganic & Medicinal Chemistry Letters*, **2006**, *16*, 4834-4838 [b] *SB alignment* [c] *LB alignment*

Table S10. Dihydropyrones Derivatives.[a]
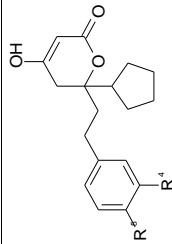
Compd.	Paper no.[a]	R^4	R^5	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
20	**21**	H	H	4.32	5.46	6.27
21	**22**	H	MeO	5.01	6.19	5.98
22	**25**	H	i-Pr	5.37	5.93	5.88
23	**26**	Me	H	4.64	5.92	5.74
24	**27**	Cl	H	5.30	6.13	5.71
25	**28**	Et	H	5.44	5.98	5.74
26	**29**	i-Pr	H	5.74	5.98	5.91
27	**31**	F	MeO	6.05	6.01	5.83
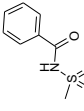
[a] *Bioorganic & Medicinal Chemistry Letters, 2006, 16, 4834-4838* [b] *SB alignment* [c] *LB alignment*

Table S11. Dihydropyrones Derivatives[a]
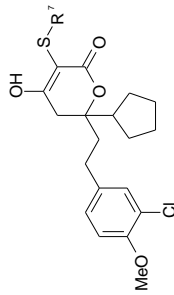
Compd.	Paper no.[a]	R[4]	R[5]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
28	**35**		28	6.25	5.54	5.97
29	**36**		29	6.89	5.55	5.92
30	**37**		30	6.55	6.25	5.85
31	**38**		31	6.85	5.85	6.00
32	**39**		32	7.42	5.43	5.75

[a] *Bioorganic & Medicinal Chemistry Letters*, **2006**, *16*, 4834-4838 [b] *SB alignment* [c] *LB alignment*
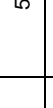
Table S12. Thiazolone derivatives[a]
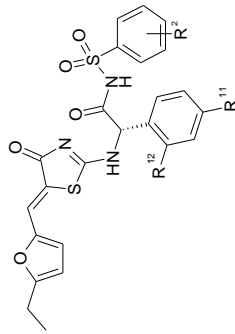
Compd.	Paper no.[a]	R[11]/R[12]	R[2]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
33	**8**	F/H	4-Me	5.03	5.51	5.57
34	**10**	F/H	3-NO2	4.82	6.15	5.3
35	**12**	F/H	3-COOH	4.92	5.55	5.17
36	**21**	H/H		4.8	5.83	5.89

[a] *Bioorganic & Medicinal Chemistry Letters*, **2007**, *17*, 1991-1995 [b] *SB alignment* [c] *LB alignment*

Table S13. Thiophene based derivatives[a]
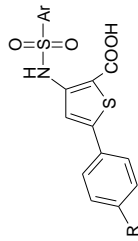
Compd	Paper no.[a]	R	Ar	exp. pIC_{50}	pred. pIC_{50} [b]	pred. pIC_{50} [c]
37	**1**	H		4.85	6.03	6.42
38	**2**	H		5.27	6.01	6.29
39	**3**	H		4.89	5.5	6.33
40	**4**	H		5.28	6.26	6.38
41	**5**	H		5.13	5.99	6.46
42	**6**	H		5.34	6.01	6.41

NS5B RESULTS ANALYSIS

Compd	Paper no.[a]	R	Ar	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
43	7	H		5.39	5.99	6.31
44	8	H		6.12	5.58	6.33
45	10			5.85	5.89	6.55
46	12	$MeSO_2$		5.74	5.99	6.48

[a] *Bioorganic & Medicinal Chemistry Letters*, **2004**, *14, 793-796* [b] *SB alignment* [c] *LB alignment*

Table S14 Thiophene-Based Derivatives[a]
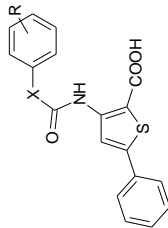
Compd	Paper no.[a]	R	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
47	**15**	2-Me	4.85	5.55	6.27
48	**16**	4-Me	4.92	6.02	6.23
49	**18**	4-Cl	4.89	5.66	6.16

[a] *Bioorganic & Medicinal Chemistry Letters*, **2004**, *14*, 793-796 [b] *SB alignment* [c] *LB alignment*

Table S15. Thiophene-Based Derivatives[a]
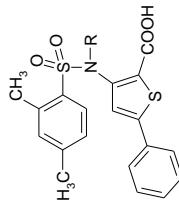
Compd	Paper no.[a]	R	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
50	**2**	Me	5.30	5.98	5.89

[a] *Bioorganic & Medicinal Chemistry Letters*, **2004**, *14, 797-800* [b] *SB alignment* [c] *LB alignment*
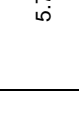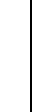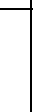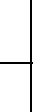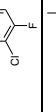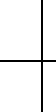
Table S16. Thiophene-Based Derivatives.[a]
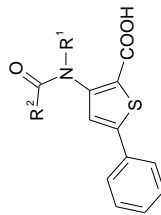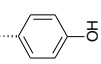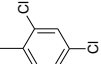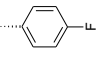
Compd.	Paper no.[a]	R[1]	R[2]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
51	**6**	Me	2-MePh	5.30	5.98	5.89
52	**9**	Me	4-CF₃Ph	4.68	5.72	6.14
53	**13**	Et	4-MePh	4.47	5.55	6.06
54	**14**	iPr	4-MePh	5.70	5.61	5.18
55	**15**	iPr	2,4-Cl₂Ph	5.52	5.85	5.96
56	**18**	iPr		5.92	5.31	5.34

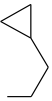[a] *Bioorganic & Medicinal Chemistry Letters*, **2004**, *14*, 797-800[b] *SB alignment* [c] *LB alignment*

Table S17. Thiazolone-Based Derivatives.[a]

Compd	Paper no.[a]	R^1	R^2	exp. pIC$_{50}$	pred. pIC$_{50}$[b]	pred. pIC$_{50}$[c]
57	**1**	4-F-phenyl	OH	5.52	6.02	5.60
58	**15**	4-F-phenyl	OH	4.30	6.34	5.78
59	**16**	4-Br-phenyl	OH	5.22	5.93	5.31
60	**17**	2,4-dichlorophenyl	OH	5.12	6.24	5.55
61	**18**	4-CF$_3$-phenyl	OH	4.68	5.98	5.51

NS5B RESULTS ANALYSIS

Compd	Paper no.[a]	R[1]	R[2]	exp. pIC$_{50}$	pred. pIC$_{50}$[b]	pred. pIC$_{50}$[c]
62	19		OH	4.85	5.71	6.02
63	20		OH	4.96	6.02	6.04
64	21		OH	4.60	5.87	5.68
65	22		OH	4.26	6.09	5.63
66	24		OH	4.62	5.99	5.66
67	25		OH	4.17	5.73	5.71
68	26		OH	4.92	5.94	5.64

NS5B RESULTS ANALYSIS

Compd	Paper no.[a]	R^1	R^2	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
69	**27**		OH	4.37	6.04	5.59
70	**28**		-NHOH	5.26	5.91	5.76
71	**29**		-NH₂	4.80	6.05	5.62

[a] *Bioorganic & Medicinal Chemistry Letters,* **2007**, *17, 63-67* [b] *SB alignment* [c] *LB alignment*

Table S18. Thiazolone-Based Derivatives.[a]

Compd.	Paper no.[a]	R	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
72	30	F-phenyl	4.36	5.40	5.82
73	32	Br-phenyl	4.78	5.73	5.47
74	33	Cl,Cl-phenyl	5.07	5.36	5.71
75	34	CH_3-phenyl	4.57	5.63	5.65

NS5B RESULTS ANALYSIS

Compd.	Paper no.[a]	R	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
76	**36**		4.23	6.01	5.69
77	**37**		4.89	6.05	5.94
78	**38**		5.05	5.92	5.73

[a] *Bioorganic & Medicinal Chemistry Letters,* **2007**, *17, 63-67* [b] *SB alignment* [c] *LB alignment*
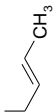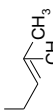
Table S19 Thiazolone-Based Derivatives.[a]
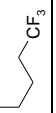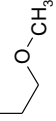
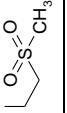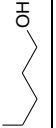| Compd.	Paper no.[a]	R	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
79	**39**	n = 0	4.72	5.81	5.66
80	**40**	n = 0 CH_3	4.85	5.50	6.28
81	**42**	n = 1	4.62	6.30	5.64
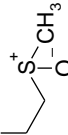
[a] Bioorganic & Medicinal Chemistry Letters, **2007**, 17, 63-67 [b] SB alignment [c] LB alignment
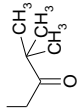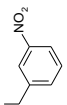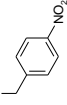
Table S20. Benzothiadiazine Derivatives. [a]
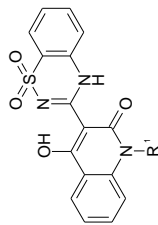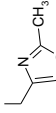
Compd.	Paper no. [a]	R[1]	exp. pIC_{50}	pred. pIC_{50} [b]	pred. pIC_{50} [c]
1	1		6.70	6.95	7.68
2	12		8.65	6.29	7.22
3	13		8.81	6.70	7.65
4	14		6.97	6.83	7.40
5	15		7.68	6.83	7.57
6	16		7.10	6.80	7.69
7	17		7.19	6.86	7.45

NS5B RESULTS ANALYSIS

Compd.	Paper no.[a]	R^1	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
8	19		6.82	6.61	7.64
9	20		6.54	6.33	7.41
10	21		7.05	7.01	7.57
11	22		6.93	7.05	7.47
12	23		6.75	6.91	7.44
13	24		7.11	6.68	7.39
14	25		7.39	6.71	7.70
15	26		7.89	6.62	7.59
16	27		7.31	7.28	7.68

Compd.	Paper no.[a]	R^1	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
17	28		6.88	6.43	7.68
18	29		6.44	6.93	7.44
19	30	CF_3	6.97	6.30	7.53
20	31	CH_3 / CF_3	7.36	6.59	7.31
21	32	Br	6.54	6.49	7.60
22	33	N	6.76	6.85	7.66
23	34	N	6.12	6.49	7.33
24	35	$O-CH_3$	6.09	6.50	7.53

NS5B RESULTS ANALYSIS

Compd.	Paper no.[a]	R[1]	exp. pIC50	pred. pIC50[b]	pred. pIC50[c]
25	36	—CH₃ / O—CH₃	7.09	6.86	7.31
26	37	(tetrahydrofuran)	6.32	7.39	7.39
27	38	(tetrahydrofuran)	6.28	6.52	7.39
28	40	S⁺—CH₃, O	8.60	6.98	7.48
29	41	O=S=O, CH₃	8.63	6.88	7.29
30	43	OH	8.55	6.72	7.24
31	44	OH	8.89	6.78	7.21
32	48	CO₂Et	8.69	6.70	7.54

Compd.	Paper no.[a]	R[1]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
33	53		6.16	7.76	7.45
34	55		7.09	6.97	7.59
35	58		6.11	6.80	7.63
36	60		7.09	6.35	7.38
37	61		8.61	5.47	7.65
38	63		7.02	6.83	7.80
39	64		7.44	6.78	7.63

NS5B RESULTS ANALYSIS

Compd.	Paper no.[a]	R[1]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
40	**65**		8.57	6.91	7.57
41	**66**		6.50	6.95	7.51
42	**67**		7.21	7.02	7.44
43	**71**		6.52	6.71	7.35

[a] *Journal of Medicinal Chemistry*, **2006**, *49*, 971-983 [b] *SB alignment* [c] *LB alignment*

Table S21. Benzothiadiazine Derivatives.[a]
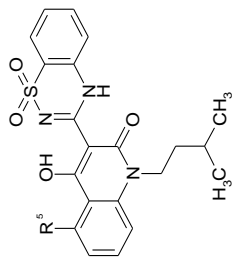
Compd.	Paper no.[a]	R^5	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
44	**72**	F	6.47	7.02	7.72
45	**74**	Br	6.98	7.28	7.44
46	**75**	CH_3	6.69	6.84	7.66
47	**76**	Ph	6.84	6.79	7.61
48	**77**	OH	7.33	6.80	7.42
49	**78**	NO_2	6.47	5.43	7.80

[a] Journal of Medicinal Chemistry, **2006**, 49, 971-983 [b] SB alignment [c] LB alignment

Table S22. Benzothiadiazine Derivatives.[a].
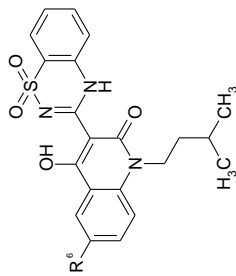
Compd.	Paper no.[a]	R^6	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
50	79	F	7.70	6.99	7.59
51	80	Cl	7.37	6.86	7.42
52	81	Br	6.69	6.64	7.75
53	83	CH_3	7.33	6.54	7.82
54	84	NO_2	6.34	6.01	7.64
55	85	NH_2	7.21	6.55	7.81
56	86	OH	6.70	6.73	7.74
57	87	OCH_3	7.64	6.69	7.39
58	89	CO_2H	8.99	5.63	7.35
59	91	OCH_2CN	7.47	6.62	7.46
60	92	$O(CH_2)_3CN$	8.36	6.59	7.39
61	93	$O(CH_2)_2OCH_3$	8.76	7.44	7.66
62	94	OCH_2CONH_2	7.62	7.02	7.92
63	95	$NHCH_2CO_2H$	7.68	6.67	7.63

NS5B RESULTS ANALYSIS

Compd.	Paper no.[a]	R[6]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
64	96	$NHCH_2COEt$	6.94	6.53	7.83
65	97a	$NH(CH_2)_2OH$	7.32	5.58	7.71
66	99	$NHCOCH_2CH(CH_3)_2$	8.19	6.07	7.78
67	100	$NHCOCH_2N(CH_3)_2$	8.24	6.99	7.63
68	101	NHCO-cyclopentyl	6.01	6.93	7.74
69	102	NHCOPh	8.37	5.95	7.69

[a] Journal of Medicinal Chemistry, 2006, 49, 971-983 [b] SB alignment [c] LB alignment
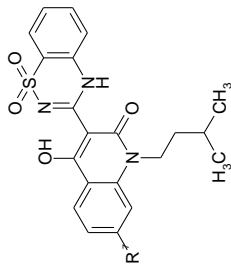
Compd.	Paper no.[a]	R^7	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
70	103	F	6.26	6.47	7.51

[a] Journal of Medicinal Chemistry, 2006, 49, 971-983 [b] SB alignment [c] LB alignment
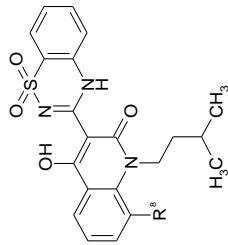
Compd.	Paper no.[a]	R[8]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
71	110	F	7.05	6.63	7.43

[a] Journal of Medicinal Chemistry, **2006**, 49, 971-983 [b] SB alignment [c] LB alignment

NS5B RESULTS ANALYSIS
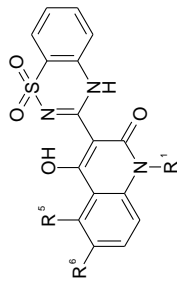
Compd.	Paper no.[a]	R[1]	R[5]	R[6]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
72	129	(cyclopropylmethyl)	H	NH_2	7.49	6.55	7.61

[a] Journal of Medicinal Chemistry, **2006**, 49, 971-983 [b] SB alignment [c] LB alignment

Table S23. Acrylic Acid Derivatives.[a]
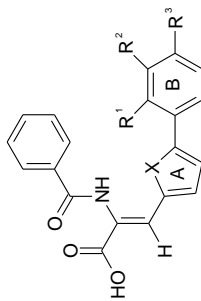
Compd.	Paper no. [a]	X	R[1]	R[2]	R[3]	exp. pIC_{50}	pred. pIC_{50} [b]	pred. pIC_{50} [c]
73	1	O	H	H	H	5.17	5.82	6.79
74	10	S	H	H	H	5.82	5.70	6.86
75	32	S	Cl	H	H	6.18	5.81	6.94
76	33	O	Cl	H	H	5.89	5.62	6.74
77	34	S	Me	H	H	5.85	5.70	6.90
78	35	O	Me	H	H	5.52	5.33	6.83
79	36	S	F	H	H	5.59	5.74	6.93
80	37	O	F	H	H	5.51	5.50	6.15
81	39	O	OCF_3	H	H	5.40	5.25	6.78
82	40	O	H	H	Cl	5.34	5.54	7.00
83	41	O	CF_3	H	H	5.25	5.58	6.70
84	42	O	H	Cl	H	5.22	5.66	6.93
85	43	O	NO_2	H	H	5.16	5.42	6.64
86	45	O	H	H	F	4.92	5.76	6.74

NS5B RESULTS ANALYSIS

Compd.	Paper no.[a]	X	R[1]	R[2]	R[3]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
87	46	O	H	H	CF_3	4.89	6.23	6.76
88	47	O	OMe	H	H	4.70	5.19	7.00
89	49	O	CH_2OH	H	OCF_3	4.66	6.16	7.09
90	50	O	H	NO_2	H	4.60	5.72	6.78
91	51	O	H	H	SMe	4.57	5.71	6.97
92	52	O	H	OCF_3	H	4.18	5.56	6.87

[a] Bioorganic & Medicinal Chemistry Letters, **2005**, 15, 2481-2486 [b] SB alignment [c] LB alignment

Table S24. Acrylic Acid Derivatives.[a]
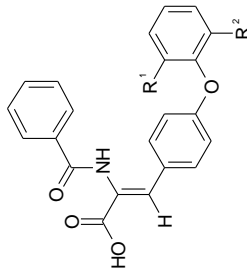
Compd.	Paper no.[a]	R[1]	R[2]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
93	31	Cl	H	6.70	5.78	6.98
94	60	Me	H	6.24	6.32	7.04
95	62	SMe	H	5.74	5.83	7.06
96	63	F	H	5.16	6.50	6.95
97	64	Et	H	5.03	6.81	7.16
98	65	OMe	H	4.82	6.64	6.99
99	66	H	H	4.77	5.83	7.04
100	67	i-Pr	H	4.55	5.77	6.99

NS5B RESULTS ANALYSIS

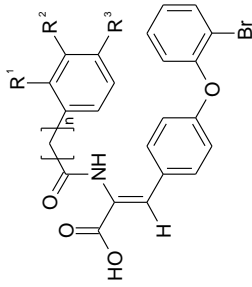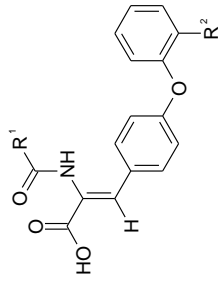| Compd.	Paper no.[a]	R[1]	R[2]	exp. pIC$_{50}$	pred. pIC$_{50}$[b]	pred. pIC$_{50}$[c]
101	68	Ph	H	4.32	5.69	7.04
102	69	O-iPr	H	4.36	6.43	7.01
103	70	CN	H	4.33	5.52	7.06
104	71	NO$_2$	H	4.14	5.61	7.09
105	73	Br	F	7.52	5.71	7.08
106	74	Cl	F	7.15	5.66	7.12
107	75	Cl	Me	6.70	5.68	7.17
108	76	Cl	NH$_2$	6.51	6.14	7.05
109	77	Cl	NO$_2$	5.26	5.72	7.03

[a] Bioorganic & Medicinal Chemistry Letters, **2005**, *15*, 2481-2486 [b] SB alignment [c] LB alignment

Table S25. Acrylic Acid Derivatives.[a]

Compd.	Paper no.[a]	n	R^1	R^2	R^3	exp. pIC$_{50}$	pred. pIC$_{50}$[b]	pred. pIC$_{50}$[c]
110	**2**	0	H	H	H	7.00	6.45	6.95
111	**29**	0	OH	H	H	7.10	5.91	7.04
112	**30**	0	H	F	H	7.05	6.02	7.01
113	**31**	0	H	H	Cl	6.96	5.88	6.98
114	**32**	0	H	H	Br	6.92	5.71	6.98
115	**33**	0	F	H	H	6.74	5.66	7.01
116	**34**	0	F	H	F	6.64	5.66	7.17
117	**35**	0	H	F	F	6.64	5.62	7.18
118	**36**	0	H	H	Me	6.60	6.44	6.86
119	**38**	1	H	H	Cl	6.03	5.57	6.85
120	**39**	1	H	H	I	6.00	6.50	6.81

NS5B RESULTS ANALYSIS

Compd.	Paper no.[a]	n	R[1]	R[2]	R[3]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
121	40	1	H	H	CF_3	6.00	6.36	6.81
122	41	1	H	H	F	5.66	5.91	6.89
123	42	2	H	H	CF_3	6.43	5.95	7.11
124	43	2	H	H	Cl	6.30	6.26	6.81
125	44	2	H	H	F	6.22	6.19	6.79
126	45	2	H	H	Me	5.59	5.81	7.07

[a] *Bioorganic & Medicinal Chemistry Letters*, **2005**, *15*, 2812-2818 [b] *SB alignment* [c] *LB alignment*

Table S26. Acrylic Acid Derivatives.[a]

Compd.	Paper no.[a]	R^1	R^2	exp. pIC$_{50}$	pred. pIC$_{50}$[b]	pred. pIC$_{50}$[c]
127	47	Cyclohexyl	Br	6.70	6.58	6.89
128	48		Br	7.00	5.60	7.02
129	50		Br	6.07	6.11	6.98
130	52		Br	5.66	5.88	7.01
131	53		Br	4.89	6.02	6.86
132	54	OH	I	5.70	5.73	6.85

NS5B RESULTS ANALYSIS

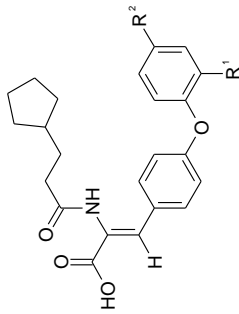| Compd.	Paper no.[a]	R¹	R²	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
133	55	Me‒⟨OH	—	5.32	6.15	6.87
134	56	Ph‒⟨OH	—	5.51	5.87	7.04
135	57	⟋⟍⟋CO_2H	—	6.10	5.58	6.84
136	58	⟋⟍⟋CO_2H	—	6.08	5.39	7.03

[a] *Bioorganic & Medicinal Chemistry Letters*, **2005**, *15*, 2812-2818 [b] *SB alignment* [c] *LB alignment*

Table S27. Acrylic Acid Derivatives.[a]

Compd.	Paper no.[a]	R^1	R^2	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
137	59	Br	H	5.22	6.46	7.16
138	60	CF$_3$	H	5.20	5.81	6.94
139	61	Et	H	5.15	6.14	6.94
140	62	Br	F	4.82	5.42	7.08
141	63	CF$_3$	NH$_2$	4.85	6.18	6.99

[a] *Bioorganic & Medicinal Chemistry Letters*, **2005**, *15*, 2812-2818 [b] *SB alignment* [c] *LB alignment*

Table S28. Acrylic Acid Derivatives.[a]
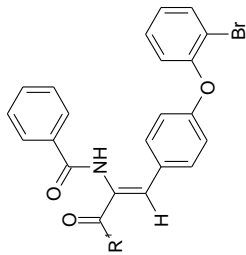
Compd.	Paper no.[a]	R[1]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
142	25	NH$_2$	6.30	6.87	7.06
143	26	NHNH$_2$	6.00	6.47	7.08
144	27	NHOMe	6.10	6.58	6.95
145	64	NH(CH$_2$)$_2$OH	5.57	6.60	7.18
146	65	NHCH$_2$CO$_2$H	7.22	5.70	6.84
147	68	NH[(4-OH)-Ph]	4.92	6.17	7.14
148	69	NH[(3-OH)-Ph]	5.72	6.42	6.97
149	70	NH[(2-OH)-Ph]	4.82	6.53	7.13
150	71	NH[(2-CO$_2$H)-Ph]	6.40	5.42	7.49

[a] Bioorganic & Medicinal Chemistry Letters, **2005**, 15, 2812-2818 [b] SB alignment [c] LB alignment
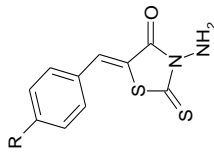
Table S29. Benzylidene Derivatives.[a]

Compd.	Paper no.[a]	R	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
151	**2e**	Cl	4.80	6.59	6.42
152	**2f**	CN	5.15	6.23	6.51
153	**2g**	F	5.15	6.56	6.39
154	**2h**	NO_2	6.00	5.80	6.04

[a] Journal of Medicinal Chemistry, **2006**, 49, 1034-1046 [b] SB alignment [c] LB alignment

Table S30. Benzylidene Derivatives.[a]
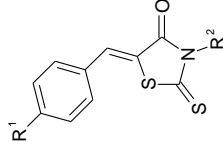
Compd.	Paper no.[a]	R^1	R^2	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
155	1	Cl	NHSO$_2$Ph	5.82	7.02	5.94
156	2m	NO$_2$	CH$_3$	4.52	5.82	6.34
157	2n	NO$_2$	H	4.66	5.84	6.62
158	5k	NO$_2$	NH$_2$SO$_2$Ph	5.70	6.34	6.06

[a] *Journal of Medicinal Chemistry*, **2006**, *49*, 1034-1046 [b] *SB alignment* [c] *LB alignment*
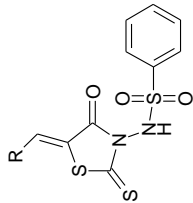
Table S31. Benzylidene Derivatives.[a]
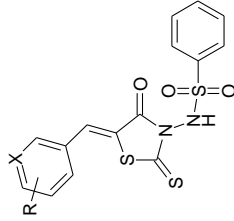
Compd.	Paper no.[a]	R	X	exp. pIC_{50}	pred pIC_{50}[b]	pred. pIC_{50}[c]
159	**5a**	H	C	4.52	6.63	5.95
160	**5b**	H	N	4.82	6.10	6.07
161	**5c**	4-CH$_3$	C	5.00	6.82	5.99
162	**5d**	3- CH$_3$	C	4.60	6.26	5.83
163	**5f**	3-OCH$_3$	C	5.30	6.27	7.48
164	**5g**	3-OCH$_2$O-4	C	4.82	6.12	6.04
165	**5i**	4-F	C	6.00	6.75	5.83
166	**5j**	2-F	C	5.19	6.60	5.88
167	**5m**	4-CF$_3$	C	6.00	6.11	5.88
168	**5n**	3-F, 4-Cl	C	6.00	5.97	5.95
169	**5o**	3,4-di-Cl	C	6.70	6.73	5.96
170	**5p**	3,4-di-Br	C	6.70	6.81	5.89
171	**5q**	2,4-di-F	C	5.52	6.42	5.92
172	**5r**	2,5-di-F	C	5.77	6.80	5.92

NS5B RESULTS ANALYSIS

Compd.	Paper no.[a]	R	X	exp. pIC_{50}	pred pIC_{50}[b]	pred. pIC_{50}[c]
173	5s	2,3,4-tri-F	C	5.72	6.05	5.80
174	5t	2,4,5-tri-F	C	6.00	6.11	5.95

[a] Journal of Medicinal Chemistry, 2006, 49, 1034-1046 [b] SB alignment [c] LB alignment

Compd.	Paper no.[a]	R	Compd.	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
175	**5w**		175	5.52	6.58	5.86

[a] Journal of Medicinal Chemistry, **2006**, 49, 1034-1046 [b] SB alignment [c] LB alignment

Table S32. Benzylidene Derivatives.[a]
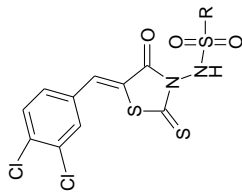
Compd	Paper no.[a]	R	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
176	6a	CH_3	5.70	6.29	7.15
177	6b	$CH_3CH_2CH_2$	6.15	6.06	6.03
178	5y	4-F-Ph	6.15	6.17	6.00
179	5z	2-F-Ph	6.70	6.55	6.02
180	5bb	3-OCH_3-Ph	6.10	6.75	6.47
181	5cc	4-OCH_3-Ph	5.22	6.18	6.31
182	5dd	4-Cl-Ph	6.30	6.40	6.24

[a] *Journal of Medicinal Chemistry*, **2006**, *49*, 1034-1046 [b] *SB alignment* [c] *LB alignment*

Table S33. Proline Sulfonamide Derivative.[a]
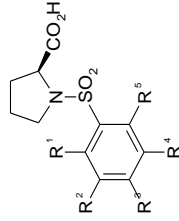
Compd.	Paper no.[a]	R^1	R^2	R^3	R^4	R^5	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
183	**18**	NH_2	H	Me	Cl	H	5.23	5.75	5.47
184	**21**	OH	Cl	H	Cl	H	6.11	5.85	5.57
185	**23**	OH	Cl	H	F	H	5.68	5.57	5.52
186	**24**	OH	Br	H	Cl	H	6.59	5.61	5.57

[a] Journal of Medicinal Chemistry, **2006**, 49, 3052-3055 [b] SB alignment [c] LB alignment

Table S34. Anthranilic Acid Derivatives.[a]
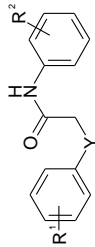
Compd.	Paper no.[a]	Y	R[1]	R[2]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
187	3e	O	4-Br	2-COOH	5.49	6.26	5.84
188	3f	O	3-Br	2-COOH	5.94	5.70	5.79
189	3i	O	4-CN	2-COOH	5.29	6.18	5.84
190	3j	O	4-F	2-COOH	5.28	5.94	5.69
191	3k	O	4-OMe	2-COOH	4.46	5.74	5.81
192	13	S	4-Cl	2-COOH	5.49	6.26	5.84
193	14n	NH	3-Br	2-COOH	5.49	6.26	5.84

[a] Journal of Medicinal Chemistry, **2007**, 50, 2108-2116 [b] SB alignment [c] LB alignment
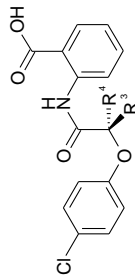
Compd	Paper no.[a]	R^3	R^4	exp. pIC$_{50}$	pred. pIC$_{50}$[b]	pred. pIC$_{50}$[c]
194	17a	H	Me	6.12	5.80	5.73
195	17b	Me	H	4.30	6.18	5.72

[a] Journal of Medicinal Chemistry, **2007**, 50, 2108-2116 [b] SB alignment [c] LB alignment

Table S35. Anthranilic Acid Derivatives.[a]
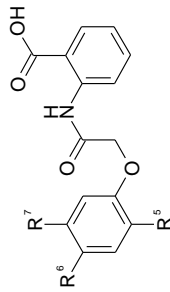
Compd.	Paper no.[a]	R[5]	R[6]	R[7]	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
196	12a	Me	Cl	H	6.24	5.93	5.87
197	12b	OMe	Cl	H	6.24	5.71	6.27
198	12c	Cl	Cl	H	6.35	6.02	6.22
199	12d	Ac	Cl	H	6.28	5.50	6.16
200	12e	Ac	H	Cl	6.07	5.67	5.71
201	12f	Ac	Me	H	5.55	6.22	7.19
202	12g	Ac	F	H	5.82	5.59	5.85
203	12h	Cl	Cl	Cl	7.52	6.22	6.37
204	12i	F	F	F	5.90	5.93	5.77
205	12j	Br	Cl	Me	6.96	6.14	5.59
206	12k	Ac	Cl	F	6.96	5.70	5.87
207	12l	Ac	F	Cl	6.64	5.23	7.24
208	12m	Ac	Cl	Me	6.85	6.23	6.45

[a] Journal of Medicinal Chemistry, **2007**, 50, 2108-2116 [b] SB alignment [c] LB alignment

Table S36. Anthranilic Acid Derivatives.[a]
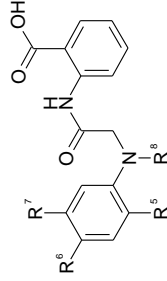
Compd	Paper no.[a]	R^5	R^6	R^7	R^8	exp. pIC_{50}	pred. pIC_{50}[b]	pred. pIC_{50}[c]
209	14a	COMe	Cl	H	H	6.64	5.67	5.41
210	14b	COEt	Cl	H	H	5.47	5.84	7.03
211	14c	COMe	Br	H	H	6.96	6.14	7.24
212	14d	COMe	Me	Me	H	7.10	5.75	5.69
213	14e	COMe	Cl	Cl	H	5.98	5.67	7.16
214	14f	COMe	F	Cl	H	7.52	6.31	5.54
215	14g	COMe	Cl	Me	H	5.13	5.95	5.57
216	14h	COMe	F	F	H	7.74	6.17	5.71
217	14j	COMe	F	Br	H	7.77	6.19	5.52
218	14k	COMe	Br	Cl	H	8.00	5.55	7.36
219	14l	COMe	Cl	Et	H	7.49	6.18	7.11
220	14m	COMe	Me	Me	H	6.68	6.19	7.10
221	14o	H	Cl	Cl	H	6.70	5.67	5.64
222	14p	H	Cl	Cl	Me	7.77	5.88	6.42
223	14q	H	F	Cl	H	6.68	5.95	5.72

[a] Journal of Medicinal Chemistry, **2007**, 50, 2108-2116 [b] SB alignment [c] LB alignment

ACKNOWLEDGMENTS

I would like to gratefully acknowledge Professor Rino Ragno who gave me the chance to do my PhD in his research group. He is not only a great scientist with a deep vision, but also a kind person. His scientific excitement inspired and supported me in the most important and difficult moments of making right decision and I am really glad to have worked in collaboration with him.

I want to give my best regards to all the people at the computational lab Rome Center for Molecular Design. It has been a pleasure to work with you all.