UNIVERSITA' DEGLI STUDI DI PARMA

Dottorato Di Ricerca
In
Progettazione E Sintesi Di Composti Biologicamente Attivi
CICIO XXVII

Development and Validation Of In Silico Tools For Efficient Library Design and Data Analysis in High Throughput Screening Campaigns

Coordinatore:
Chiar.mo Prof. Marco Mor

Tutor:
Chiar.mo Prof. Gabriele Costantino

Dottorando: Andrea Rosario Beccari

LIST OF PAPERS

Targeting the minor pocket of C5aR for the rational design of an oral allosteric inhibitor for inflammatory and neuropathic pain relief.

Moriconi A, Cunha TM, Souza GR, Lopes AH, Cunha FQ, Carneiro VL, Pinto LG, Brandolini L, Aramini A, Bizzarri C, Bianchini G, Beccari AR, Fanton M, Bruno A, Costantino G, Bertini R, Galliera E, Locati M, Ferreira SH, Teixeira MM, Allegretti M.

Proc Natl Acad Sci U S A. 2014 Nov 25;111(47):16937-42. doi: 10.1073/pnas.1417365111. Epub 2014 Nov 10.

PMID: 25385614 [PubMed - in process] Free PMC Article

Molecular mechanism and functional role of brefeldin A-mediated ADP-ribosylation of CtBP1/BARS.

Colanzi A, Grimaldi G, Catara G, Valente C, Cericola C, Liberali P, Ronci M, Lalioti VS, Bruno A, Beccari AR, Urbani A, De Flora A, Nardini M, Bolognesi M, Luini A, Corda D.

Proc Natl Acad Sci U S A. 2013 Jun 11;110(24):9794-9. doi: 10.1073/pnas.1222413110. Epub 2013 May 28.

PMID: 23716697 [PubMed - indexed for MEDLINE] Free PMC Article

State-of-the-art and dissemination of computational tools for drug-design purposes: a survey among Italian academics and industrial institutions.

Artese A, Alcaro S, Moraca F, Reina R, Ventura M, Costantino G, Beccari AR, Ortuso F.

Future Med Chem. 2013 May;5(8):907-27. doi: 10.4155/fmc.13.59.

PMID: 23682568 [PubMed - indexed for MEDLINE]

LiGen: a high performance workflow for chemistry driven de novo design.

Beccari AR, Cavazzoni C, Beato C, Costantino G.

J Chem Inf Model. 2013 Jun 24;53(6):1518-27. doi: 10.1021/ci400078g. Epub 2013 May 28.

PMID: 23617275 [PubMed - indexed for MEDLINE]

Use of experimental design to optimize docking performance: the case of LiGenDock, the docking module of LiGen, a new de novo design program.

Beato C, Beccari AR, Cavazzoni C, Lorenzi S, Costantino G.

J Chem Inf Model. 2013 Jun 24;53(6):1503-17. doi: 10.1021/ci400079k. Epub 2013 Apr 30.

PMID: 23590204 [PubMed - indexed for MEDLINE]

Development and Validation Of In Silico Tools For Efficient Library Design and Data Analysis in High Throughput Screening Campaigns

Chapter 1: Introduction

Chapter 1: Introduction

1 INTRODUCTION

1.1 RESEARCH, DRUG DISCOVERY AND DEVELOPMENT

1.1.1 The productivity of the sector

Since the human genome was sequenced ten years ago, the number of compounds in development has increased by 62% and total R&D expenditures have doubled[1] (Figure 3). Twenty nine FDA[2] approvals per year has been average over the past two decades for novel drugs classified as new molecular entities (NMEs) and biologic license applications (BLAs). Productivity per year is almost constant, after a peak in 2012 of 38 approvals, in 2013[3] they decreased to 27, demonstrating that no significant improvement has been made since 1993 (**Figure** 1)

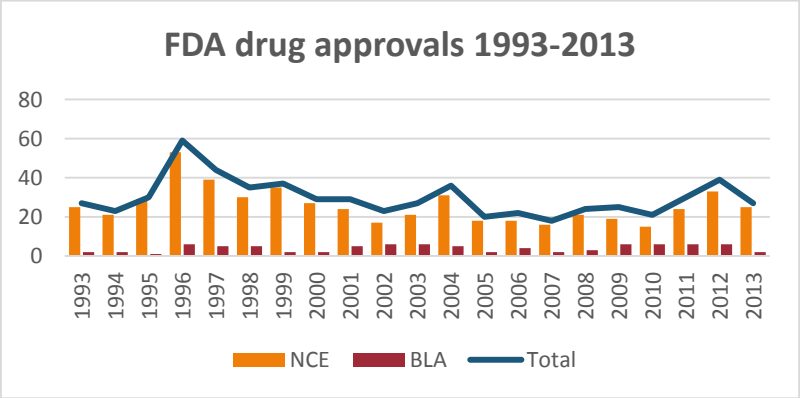


Figure 1. FDA Drug Approvals 1993-2013

The entire process to develop a new drug from the identification of a target to the registration and commercialization of new NMEs takes between 10 and 15 years[4] (**Figure** 2), with one of the lowest chances of success compared to other industrial sectors[5]. This is because of the very low probability of phase transition in clinics where the costs are higher[6] and the candidate drugs are exposed to the patients. In the last 10 years, the drug discovery process has been re-engineered and optimized obtaining a dramatic reduction in failures due to the development of poorly validated candidate drugs[7].

In 1991, the main issue for a clinical failure was drug metabolism and pharmacokinetics (DMPK),essentially, the physic-chemical properties of the molecules; in the year 2000, this problem was successfully addressed, since when the clinical safety of the candidate drugs has become the most relevant reason for a clinical failure[7].

Nowadays, low efficacy is responsible for clinical failure; the success rate is about 32% in phase II and 60% in phase III clinical trials[1]. With a global probability of success of 10%[1], the total cost of bringing a NCE into the market is 10 ten times higher than the R&D costs of the drug itself due to the 9 out of 10 failures.

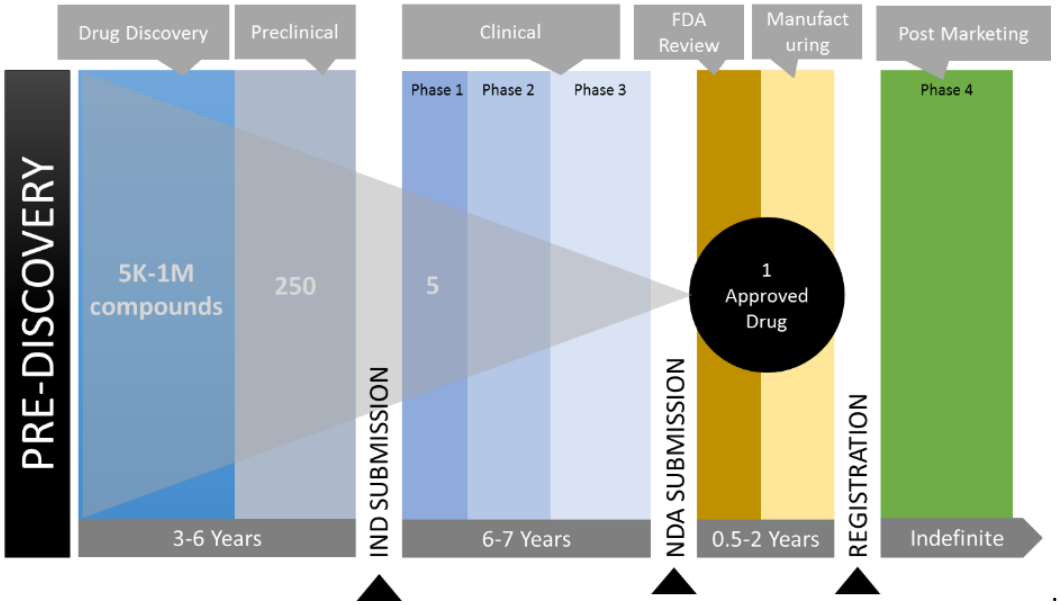


Figure 2 from Target Identification to Post Marketing (Source: Adapted from Pharmaceutical Research and Manufacturers of America, Drug Discovery and Development: Understanding the R&D Process, www.innovation.org.)

1.1.1 Economics

To maintain the same average productivity in terms of NCEs per year, global R&D costs of the pharmaceutical industry increased more than 4 fold from 31 US$ billion in 1990 to 140 US$ billion in 2008, whereas from 2008 to 2013 the spending has been almost constant with an average spending of 137 US$ billion (**Figure** 3)[8].

Over the last 6 years, the global R&D costs and productivity have remained stable meaning that the heavy cost optimization, refocusing of the research activities and massive layoffs was due to the consolidation of the transition from internal research to contract research rather than to a disinvestment in Pharma R&D.

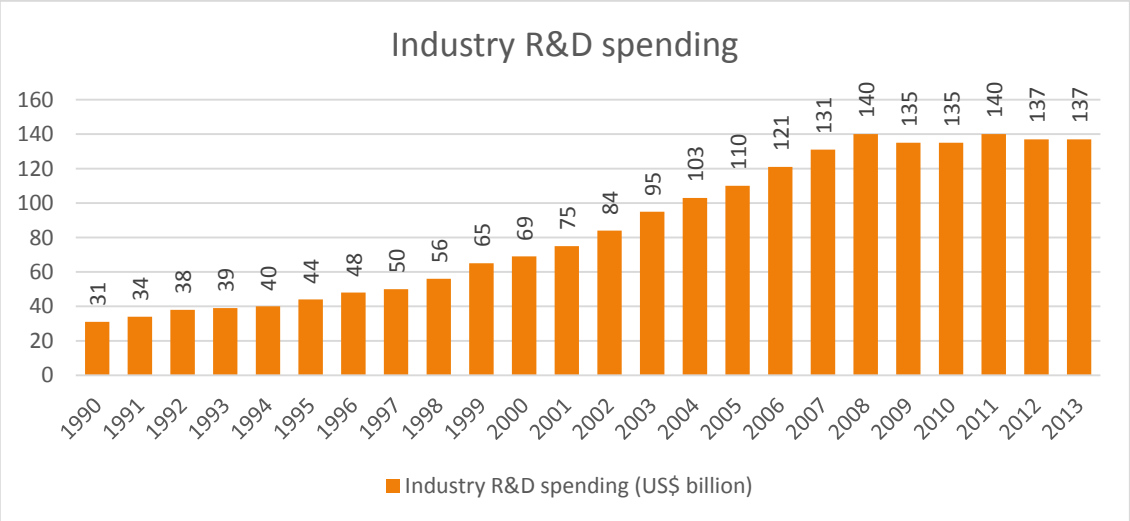


Figure 3. .Aggregate industry spending on research and development. All values inflation adjusted to 2013. (Sources: EvaluatePharma; US Food and Drug Administration (FDA); Boston Consulting Group (BCG) analysis.)

If we do not consider the registration of new NCEs as a unique criteria of productivity, extending the analysis also to the other products such as: bioengineered vaccines, combination products, enantiomers,

prodrugs, and therapeutic blood and plasma products, that can be collectively called new therapeutic drugs (NTDs)[8], the overall productivity of the sector is consistent with the allocated resources in R&D.

In fact, even if the R&D costs are estimated to be close to 2 billion for each drug reaching the market (considering the failures), the pharmaceutical market referring only to the top 25 pharmaceutical companies in 2013 is still in growth(> 2%) with a global growth in sales of more than a billion over what time period (Figure 4).

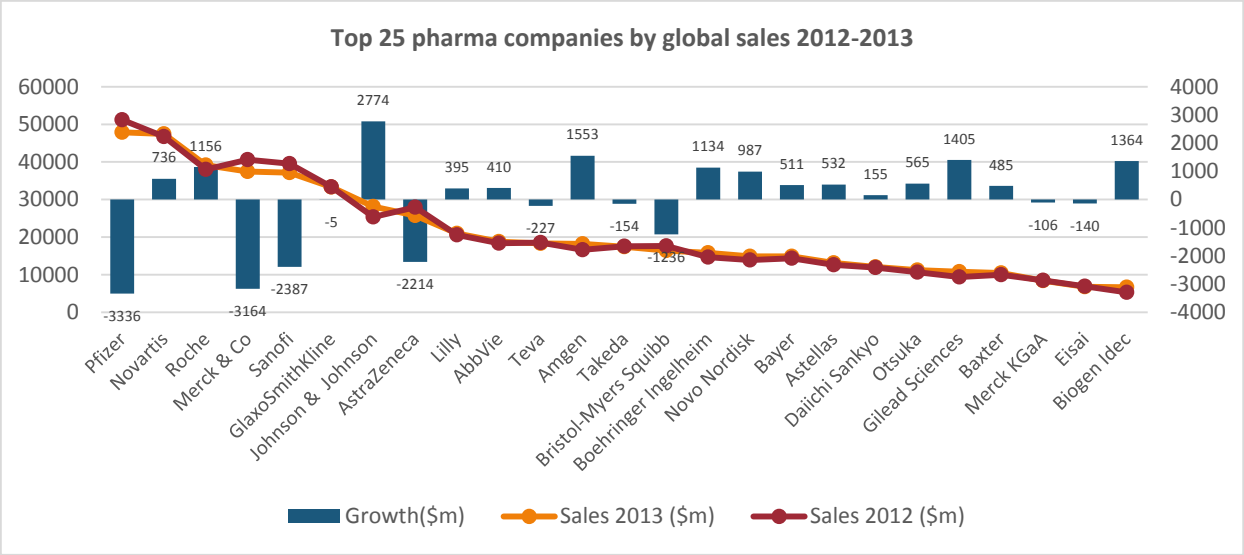


Figure 4. Top 25 pharma companies by global sales 2012-2013

If compared to other industrial sectors pharmaceuticals is one of the most profitable markets (Error! Reference source not found.) comparable to internet companies like Google and ICT leaders such as Apple Inc, with a stable outlook compared other tech and media markets.

Table 1. Profitability Analysis of key players in different industrial sectors 2014 (Source: http://www.bloomberg.com/)

Ticker	Name	Return on Capital	Profit Margin
PHARMA	Average	14.39	20.32
NOVN VX	Novartis AG	12.57	21.64
SAN FP	Sanofi	6.62	20.29
ROG VX	Roche Holding AG	30.78	26.33
GSK LN	GlaxoSmithKline PLC	19.72	20.45
AZN LN	AstraZeneca PLC	5.67	24.58
PFE US	Pfizer Inc	10.04	29.64
MRK US	Merck & Co Inc	8.07	23.72
AUTOMOT	Average	11.09	4.5
TM US	Toyota Motor Corp	6.73	
7267 JP	Honda Motor Co Ltd	5.6	
7201 JP	Nissan Motor Co Ltd	4.74	
GM US	General Motors Co	4.5	
F US	Ford Motor Co		4.5
RETAIL	Average	12.35	1.55
WMT US	Wal-Mart Stores Inc	12.98	
OIL	Average	9.12	9.72

RDSA LN	Royal Dutch Shell PLC	7.73	
TECH	Average	13.97	15.6
GOOG US	Google Inc	13.67	
YHOO US	Yahoo! Inc		35.18
FB US	Facebook Inc		27.97
AAPL US	Apple Inc	27.7	
MSFT US	Microsoft Corp	20.69	
Manuf	Average	11.89	13.83
PG US	Procter & Gamble Co/The	11.1	
KMB US	Kimberly-Clark Corp	20.47	10.56

1.1.2 How it is possible to enhance the productivity (and knowledge)?

Lack of efficacy in clinics in 68% of the phase II clinical trials highlights the biggest challenge for drug discovery and development[9]. The goal in this process is the identification of clinically validated biological targets with a proven disease modifying effect on patients and the development of effective translational models in the preclinical phase. We are, in any case still far from reaching this goal as demonstrated by the very low success rate in clinical development in oncology. With a probability Of being successful in clinics of only 7%, oncology is the therapeutic area with the lowest probability of success despite the fact that it is one of the most financed, “personalized”[10] and stratified[11] disease indication.

To make evident the effect of the attrition rate of bringing novel and effective treatment to address the unmet medical needs of society, the most urgent needs in Europe in terms of healthcare are reported, as presented by the European Commission during the Workshop on "High Performance Computing (HPC) in Health Research” in Brussels, 1 - 2 October 2014:

1. 2 million deaths in Europe due to cardiovascular diseases costing €192 billion
2. 27 million people suffering from diabetes and 120 million with rheumatic and musculoskeletal conditions with large cost implications
3. Brain disorders cost an estimated €800 billion
4. In England treatment of Alzheimer’s disease costs approximately £17 billion

Such data mean that we still have a poor understanding of the biological process and key mediators in diseases. On the other hand the R&D pharma model has to be revised in order to address the issue from a different perspective in fact, as demonstrated in the previous paragraph, the contract research model has reached the plateau of productivity.

One possible and more fruitful solution, alternative to contacts research, is the establishment of collaborations with hospitals, public research institutions and academia. This approach effectively supports the identification and clinical validation of new targets being, at the same time, as close as possible to actual patients’ conditions even in the earliest discovery phases.

Many initiatives under the buzzword “open innovation”[12] have been reported so far. Here I report the one that is, to me, a great example of a successful implementation of the open model: the Structural Genomics Consortium (SGC).

"The SGC catalyzes research in new areas of human biology and drug discovery by focusing explicitly on less well-studied areas of the human genome. The SGC accelerates research in these new areas by making all its research output available to the scientific community with no strings attached, and by creating an open collaborative network of scientists in hundreds of universities around the world and in nine global pharmaceutical companies. Together, this network of academic and industry scientists is driving a new scientific and drug discovery ecosystem whose primary aim is to advance science and is less influenced by personal, institutional or commercial gain. " << http://www.thesgc.org/about/what_is_the_sgc >>

SGC embraces the entire drug discovery and development process from the identification of novel targets to the clinical validation of the identified chemical probes. SGC has already been successful in both these activities demonstrating that radically new research models avoiding data access restriction and without any patent protection can bring real benefits to patients and nevertheless be valuable products for pharmaceutical industries.

As said before a big issue is the poor comprehension of the key biological process related to a pathology. To enhance the knowledge and then the productivity of the sector, data mining, analysis and modeling and computer simulations could be part of the solution to the problem. In the next five years, exascale computing will become a reality and will bring to the research community huge data storage facilities with almost unlimited computing power. Starting from software as a service (SaaS) with many of them already available today in the terascale era, to Platforms as a Service (PaaS) and entire Virtual research environments (VRE) the ICT/HPC the computer could be the driving force of the change.

"VREs are expected to result in more effective collaboration between researchers and higher efficiency and creativity in research as well as in higher productivity of researchers thanks to reliable and easy access to discovery, access and re-use of data. They will accelerate innovation in research via an integrated access to potentially unlimited digital research resources, tools and services across disciplines and user communities and enable researchers to process structured and qualitative data in virtual and/or ubiquitous workspaces. They will contribute to increased take-up of collaborative research and data sharing by new disciplines, research communities and institutions."
"VRE should abstract from the underlying e-infrastructures using standardized building blocks and workflows, well documented interfaces, in particular regarding APIs, and interoperable components. Over time VREs will be composed of generic services delivered by e-infrastructures and domain specific services co-developed and co-operated by researchers, technology and e-infrastructure providers, and possibly commercial vendors."
<<http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/2144-einfra-9-2015.html >>

Therefore, exascale is not only limited to the vertical application of huge computer simulations but the exascale can be reached by a pervasive use of ICT and HPC architectures from the entire community of researchers from theoretical physics to the doctor in the hospital.

To achieve this goal the Research and Innovation department of the European Commission has designed specific grants in the Horizon 2020 program from VRE to Centres of Excellence (CoE) for computing applications.

Even more relevant than the infrastructure is the possibility of having access to relevant information[13], such as Medical claims, Prescription claims, Electronic medical records (EMRs), Diagnostics, Biomarkers,

Admission, discharge, transfer, Drug orders and sales, Safety and Pharmaco-vigilance , Clinical research, Social media in combination with environmental geo referenced data such as climate, food consumption, pollutions, electromagnetic sources …

The combination of ICT/HPC infrastructures with huge collections of curated data assure the acquisition of new knowledge by the research community although this promise will be maintained only if the software for the data analytics and in silico simulations are ready to take full advantage of the present and future infrastructure, but which is currently more of a software the bottleneck.

1.2 Computer Aided Drug Discovery and Development

Computer-aided[14] Drug Discovery and Development (CADDD) approaches have been widely used in pharmaceutical research to improve the efficiency of the drug discovery and development pipeline. To identify and design small molecules as clinically effective therapeutics, various computational methods have already been evaluated as promising strategies. The original definition refers to the most established applications mainly in the medicinal chemistry area:

- the conformational analysis of molecular structure: e.g., molecular dynamics
- the characterization of drug–target interactions: e.g., molecular docking
- the identification of most relevant features shared by a set of active ligands: pharmacophore analysis
- the analysis of the chemical space and chemical library design: cheminformatics
- the assessment and optimization of drug activity using quantitative and qualitative structure-activity relationships: e.g., Q(q)SARs.

The amount and variety of biological data now available, together with techniques developed so far have enabled research in Bioinformatics to move beyond the study of individual biological components (genes, proteins etc), albeit in a genome-wide context, to attempt to study how individual parts cooperate in their operation. Bioinformatics has now moved closer to the area of Systems Biology which seeks to integrate biological data in an attempt to understand how biological systems function. By studying the relationships and interactions between various parts of a biological system, it is hoped that an understandable model of the whole system can be developed.

Bioinformatics and Systems Biology are now fully integrated in the drug discovery and development process.

Other CADDD disciplines have an ever-increasing role in the R&D process.

Pharmacometrics[15] or Quantitative Pharmacology is defined as the science that quantifies drug, disease and trial information to aid efficient drug development, regulatory decisions and rational drug treatment in patients. Pharmacometrics uses models based on systems pharmacology, physiology and disease for quantitative analysis of interactions between drugs and patients. This involves pharmacokinetics, pharmacodynamics and disease progression with a focus on populations and variability. In the field of phamacometrics for example, compartmental models for predicting oral drug absorption[16] including compartmental absorption and transit model; Grass model; the gastrointestinal transit absorption model; advanced compartmental absorption and transit model; and advanced dissolution, absorption, and metabolism model have demonstrated greatly improved predictive performance if compared with mid-1990s QSAR models based on physic-chemical descriptors.

One example of the frontier of the contemporary biomedicine from a computational perspective is the Virtual Physiological Human Project[17]. The VPH is a framework of methods and technologies that once established will make it possible to investigate the human body as a whole. VPH is a Large-scale research initiative started in 2005 with > 200 million Euro and involving more than 2000 researchers in Europe. The final goal of the project is to allow the possibility to perform *In silico* clinical trials.

In silico clinical trials will use computer simulation in the development and in the assessment of new drugs, new medical devices, new health technologies. *In silico* clinical trials will use patient-specific computer modelling and simulation applied in the research and development of new biomedical products, as well as in their evaluation and assessment, both clinical and pre-clinical.

Therefore, the question is, in the highlighted futuristic scenario, Is there still interest in the further development of traditional structure-based and ligand based algorithms and tools?

The answer is yes, and for many reasons. The most relevant ones are:

1. Traditional molecular level computer simulations are complementary to the compartmental models.
2. Even if well established, the algorithms (for example the scoring functions of the molecular docking tools) are far from being able to produce accurate ligand-receptor affinities predictions.
3. Most of the available software is not designed for real HPC simulation so it will be not possible to take full advantages of the exascale architecture.

More efficient tools in early drug discovery have demonstrated their usefulness in the reduction of the attrition rate[18] in the preclinical phase by enriching for example the number of hits in High Throughput Screening (HTS) Campaigns or in the development of robust SARs.

The compelling need for efficient computer simulation tools enabled for HPC infrastructures is demonstrated by the creation worldwide only this year (2014) of 3 HPC centers devoted to molecular simulations. Press releases are reported:

1. “Fujitsu Provides TC Cloud Service to the **University of Tokyo's RCAST** for In Silico Drug Discovery Research”
2. **Fujitsu HPC Competency Center** “envisages that demand for HPC in India will stem from application areas such as advanced drug discovery, … “
3. **Lawrence Livermore National Laboratory** A New Model for Pharma «Pharmaceutical companies need a faster and more accurate way to identify promising drug compounds and evaluate the efficacy and safety of new drugs.»
4. **Novartis Institutes for Biomedical Research** Taps Cloud HPC for Faster Drug Discovery & Better Science

The Novartis Institutes for Biomedical Research press release, clearly state that real HPC Virtual Screening is not jet integrated in the drug discovery process and, even in big pharmaceutical companies, it is so episodic and exceptional that a press release is needed.

1.3 References

1. Hay, M.; Thomas, D. W.; Craighead, J. L.; Economides, C.; Rosenthal, J., Clinical development success rates for investigational drugs. *Nature biotechnology* **2014,** *32* (1), 40-51.

2. Mullard, A., 2012 FDA drug approvals. *Nature reviews. Drug discovery* **2013,** *12* (2), 87-90.
3. Mullard, A., 2013 FDA drug approvals. *Nature reviews. Drug discovery* **2014,** *13* (2), 85-9.
4. (a) DiMasi, J. A.; Feldman, L.; Seckler, A.; Wilson, A., Trends in risks associated with new drug development: success rates for investigational drugs. *Clinical pharmacology and therapeutics* **2010,** *87* (3), 272-7; (b) Kaitin, K. I.; DiMasi, J. A., Pharmaceutical innovation in the 21st century: new drug approvals in the first decade, 2000-2009. *Clinical pharmacology and therapeutics* **2011,** *89* (2), 183-8.
5. (a) Brown, D.; Superti-Furga, G., Rediscovering the sweet spot in drug discovery. *Drug discovery today* **2003,** *8* (23), 1067-77; (b) Garnier, J. P., Rebuilding the R&D engine in big pharma. *Harvard business review* **2008,** *86* (5), 68-70, 72-6, 128; (c) Pammolli, F.; Magazzini, L.; Riccaboni, M., The productivity crisis in pharmaceutical R&D. *Nature reviews. Drug discovery* **2011,** *10* (6), 428-38.
6. Bunnage, M. E., Getting pharmaceutical R&D back on target. *Nature chemical biology* **2011,** *7* (6), 335-9.
7. Khanna, I., Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug discovery today* **2012,** *17* (19-20), 1088-102.
8. Schulze, U.; Baedeker, M.; Chen, Y. T.; Greber, D., R&D productivity: on the comeback trail. *Nature reviews. Drug discovery* **2014,** *13* (5), 331-2.
9. Arrowsmith, J.; Miller, P., Trial watch: phase II and phase III attrition rates 2011-2012. *Nature reviews. Drug discovery* **2013,** *12* (8), 569.
10. (a) Bartlett, R.; Everett, W.; Lim, S.; G, N.; Loizidou, M.; Jell, G.; Tan, A.; Seifalian, A. M., Personalized in vitro cancer modeling - fantasy or reality? *Translational oncology* **2014,** *7* (6), 657-64; (b) Nofziger, C.; Papaluca, M.; Terzic, A.; Waldman, S.; Paulmichl, M., Policies to aid the adoption of personalized medicine. *Nature reviews. Drug discovery* **2014,** *13* (3), 159-60.
11. (a) Chataway, J.; Fry, C.; Marjanovic, S.; Yaqub, O., Public-private collaborations and partnerships in stratified medicine: making sense of new interactions. *New biotechnology* **2012,** *29* (6), 732-40; (b) Foot, E., Stratified medicine: making it happen. *Drug discovery today* **2011,** *16* (19-20), 848-9; (c) Trusheim, M. R.; Berndt, E. R.; Douglas, F. L., Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nature reviews. Drug discovery* **2007,** *6* (4), 287-93.
12. (a) Schuhmacher, A.; Germann, P. G.; Trill, H.; Gassmann, O., Models for open innovation in the pharmaceutical industry. *Drug discovery today* **2013,** *18* (23-24), 1133-7; (b) Wang, L.; Plump, A.; Ringel, M., Racing to define pharmaceutical R&D external innovation models. *Drug discovery today* **2014**.
13. Szlezak, N.; Evers, M.; Wang, J.; Perez, L., The role of big data and advanced analytics in drug discovery, development, and commercialization. *Clinical pharmacology and therapeutics* **2014,** *95* (5), 492-5.
14. (a) Kapetanovic, I. M., Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. *Chemico-biological interactions* **2008,** *171* (2), 165-76; (b) Zhang, S., Computer-aided drug discovery and development. *Methods in molecular biology* **2011,** *716*, 23-38.
15. Barrett, J. S.; Fossler, M. J.; Cadieu, K. D.; Gastonguay, M. R., Pharmacometrics: a multidisciplinary field to facilitate critical thinking in drug development and translational research settings. *Journal of clinical pharmacology* **2008,** *48* (5), 632-49.
16. Huang, W.; Lee, S. L.; Yu, L. X., Mechanistic approaches to predicting oral drug absorption. *The AAPS journal* **2009,** *11* (2), 217-24.
17. Viceconti, M., Biomechanics-based in silico medicine: The manifesto of a new science. *Journal of biomechanics* **2014**.
18. Gasteiger, J.; Engel, T., *Chemoinformatics : a textbook*. Wiley-VCH: Weinheim, 2003; p xxx, 649 p.

Development and Validation Of In Silico Tools For Efficient Library Design and Data Analysis in High Throughput Screening Campaigns

Chapter 2: Aim Of The Here Reported Ph.D. Studies

Chapter 2: Aim Of The Here Reported Ph.D. Studies

2 AIM OF THE HERE REPORTED PH.D. STUDIES

In the scenario given in the introduction (chapter 1), data analytics and computer simulations already form an integral part of the entire drug discovery and development process and will dramatically increase their relevance in the future as has already happened in other advanced, from a computational point of view, sectors such as material science and the aerospace/automotive fields.

My PhD project findings have their major application in the early phase of the drug discovery process, in particular we have developed and validated two computational tools (Molecular Assembles and LiGen) to support the hit finding and the hit to lead phases.

I have reported here novel methods to first design chemical libraries optimized for HTS and then profile them for a specific target receptor or enzyme. I also analyzed the generated bio-chemical data in order to obtain robust SARs and to select the most promising hits for the follow up. The described methods support the iterative process of validated hit series optimization up to the identification of a lead.

In chapter 3, Ligand generator (LiGen), a *de novo* tool for structure based virtual screening, is presented. The development of LiGen is a project based on a collaboration among Dompé Farmaceutici SpA, CINECA and the University of Parma. In this multidisciplinary group, the integration of different skills has allowed the development, from scratch, of a virtual screening tool, able to compete in terms of performance with long standing, well-established molecular docking tools such as Glide, Autodock and PLANTS.

LiGen, using a novel docking algorithm, is able to perform ligand flexible docking without performing a conformational sampling. LiGen also has other distinctive features with respect to other molecular docking programs:

- LiGen uses the inverse pharmacophore derived from the binding site to identify the putative bioactive conformation of the molecules, thus avoiding the evaluation of molecular conformations which do not match the key features of the binding site.
- LiGen implemenst a *de novo* molecule builder based on the accurate definition of chemical rules taking account of building block (reagents) reactivity.
- LiGen is natively a multi-platform C++ portable code designed for HPC applications and optimized for the most recent hardware architectures like the Xeon Phi Accelerators.

Chapter 3 also reports the further development and optimization of the software starting from the results obtained in the first optimization step performed to validate the software and to derive the default parameters.

In chapter 4, the application of LiGen in the discovery and optimization of novel inhibitors of the complement factor 5 receptor (C5aR) is reported. Briefly, the C5a anaphylatoxin acting on its cognate G protein-coupled receptor C5aR is a potent pronociceptive mediator in several models of inflammatory and neuropathic pain. Although there has long been interest in the identification of C5aR inhibitors, their development has been complicated, as is the case with many peptidomimetic drugs, mostly due to the poor drug-like properties of these molecules. Herein, we report the *de novo* design of a potent and selective C5aR noncompetitive allosteric inhibitor, DF2593A. DF2593A design was guided by the hypothesis that an allosteric site, the “minor pocket”, previously characterized in CXCR1 and CXCR2, could be functionally conserved in the GPCR class.DF2593A potently inhibited C5a-induced migration of human and rodent neutrophils *in vitro*. Moreover, oral administration of DF2593A effectively reduced mechanical

hyperalgesia in several models of acute and chronic inflammatory and neuropathic pain *in vivo*, without any apparent side effects.

Chapter 5 describes another tool: Molecular Assemblies (MA), a novel metrics based on a hierarchical representation of the molecule based on different representations of the scaffold of the molecule and pruning rules.

The algorithm used by MA, defining *a priori* a metrics (a set of rules), creates a representation of the chemical structure through hierarchical decomposition of the scaffold in fragments, in a pathway invariant way (this feature is novel with respect to the other algorithms reported in literature) (Figure 5**Error! Reference source not found.**).

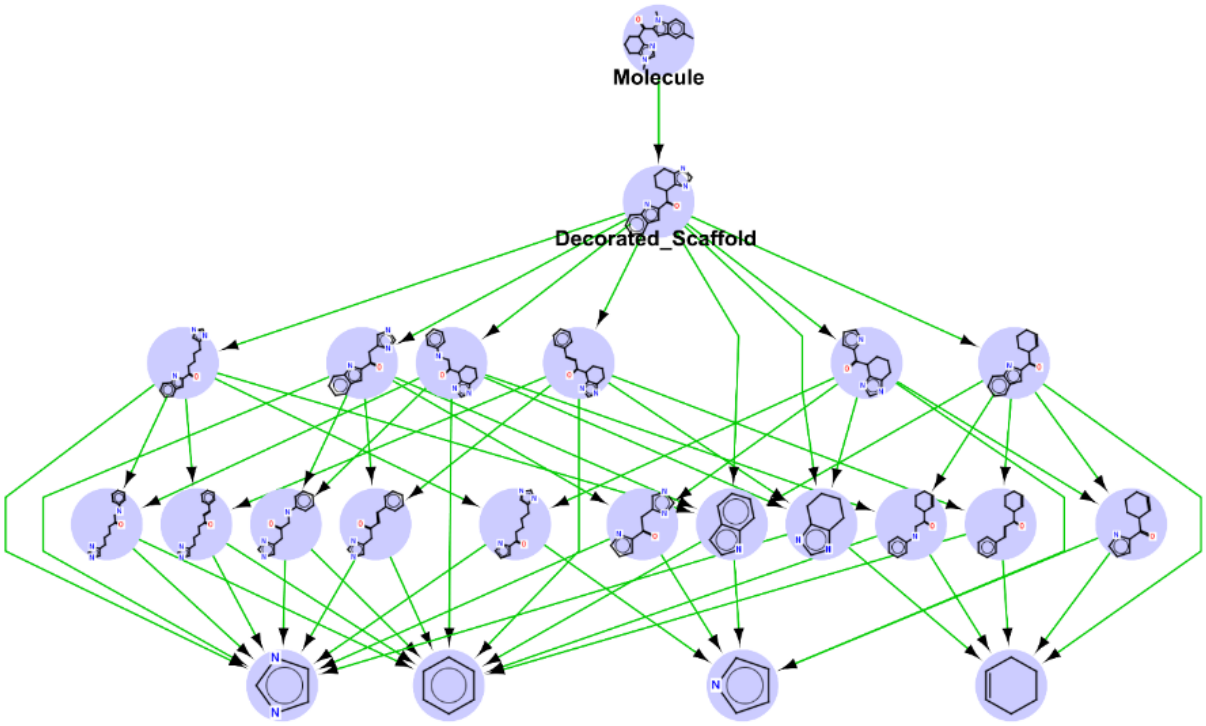


Figure 5. Cystoscape plot of the pathway independent fragmentation of one of the nine representations of the reference molecule

Such structure decomposition is applied to nine hierarchical representation of the scaffold of the reference molecule, differing for the content of structural information: atom typing and bond order (this feature is novel with respect to the other algorithms reported in literature) (Figure 6).

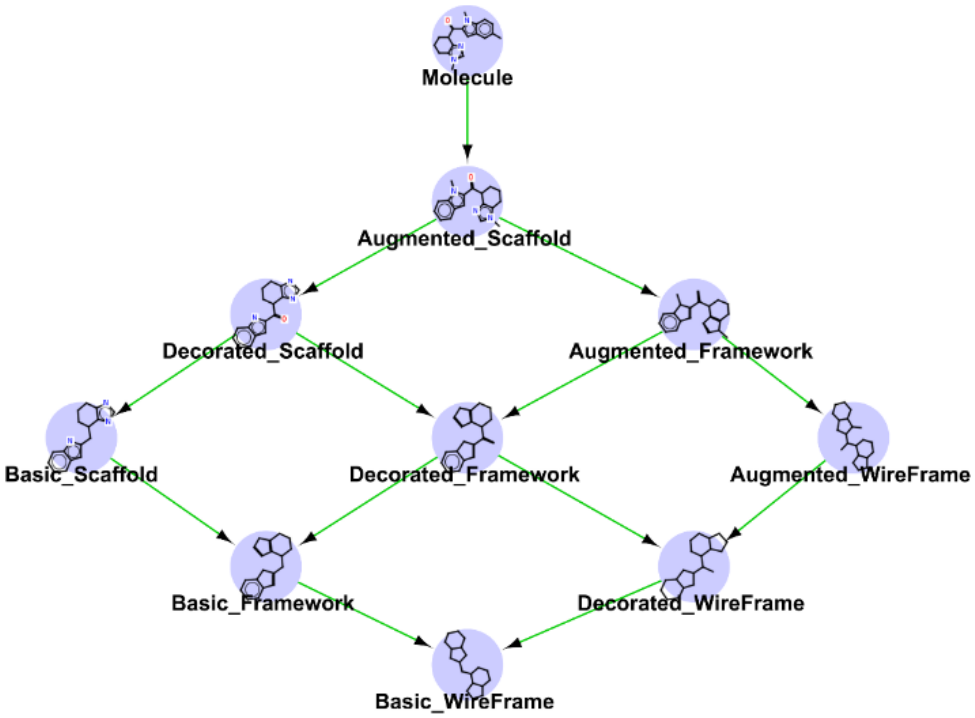


Figure 6. Cystoscape plot of the nine hierarchical representations of the reference molecule

The algorithm (metrics) generates a multi-dimensional hierarchical representation of the molecule.

This descriptor applied to a library of compounds is able to extract structural (molecule having the same scaffold, wireframe or framework) and sub structural (molecule having the same fragments in common) relations among all the molecules.

At least, this method generates relations among molecules based on identities (scaffolds or fragments). Such an approach produces a unique representation of the reference chemical space not biased by the threshold used to define the similarity cut-off between two molecules. This is in contrast to other methods which generate representations based in similarities.

MA procedure, retrieving all scaffold representation, fragments and fragmentation's patterns (according to the predefined rules) from a molecule, creates a molecular descriptor useful for several cheminformatics applications:

- Visualization of the chemical space. The scaffold relations (Figure 7) and the fragmentation patterns can be plotted using a network representation. The obtained graphs are useful depictions of the chemical space highlighting the relations that occur among the molecule in a two dimensional space.
- Clustering of the chemical space. The relations among the molecules are based on identities. This means that the scaffold representations and their fragments can be used as a hierarchical clustering method. This descriptor produces clusters that are independent from the number and similarity among closest neighbors because belonging to a cluster is a property of the single molecule (Figure 8). This intrinsic feature makes the scaffold based clustering much faster than other methods in producing "stable" clusters in fact, adding and removing molecules increases and decreases the number of clusters while adding or removing relations among the clusters. However these changes do not affect the cluster number and the relation of the other molecules in dataset.

- Generate scaffold-based fingerprints. The descriptor can be used as a fingerprint of the molecule and to generate a similarity index able to compare single molecules or also to compare the diversity of two libraries as a whole.

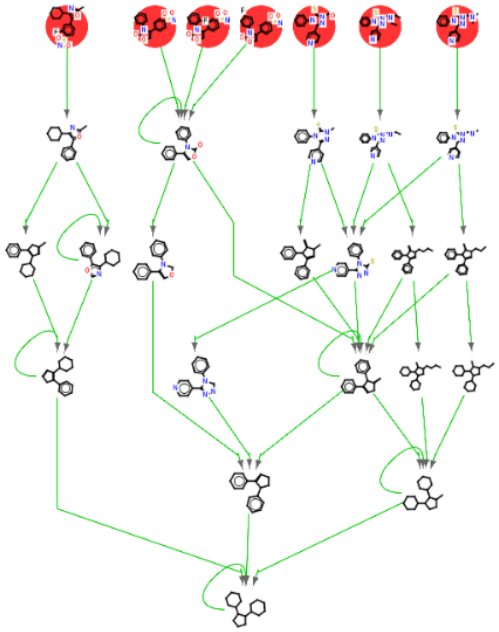


Figure 7. Cystoscape Plot of the network of structural relationships generate by the Molecular Assemblies descriptor

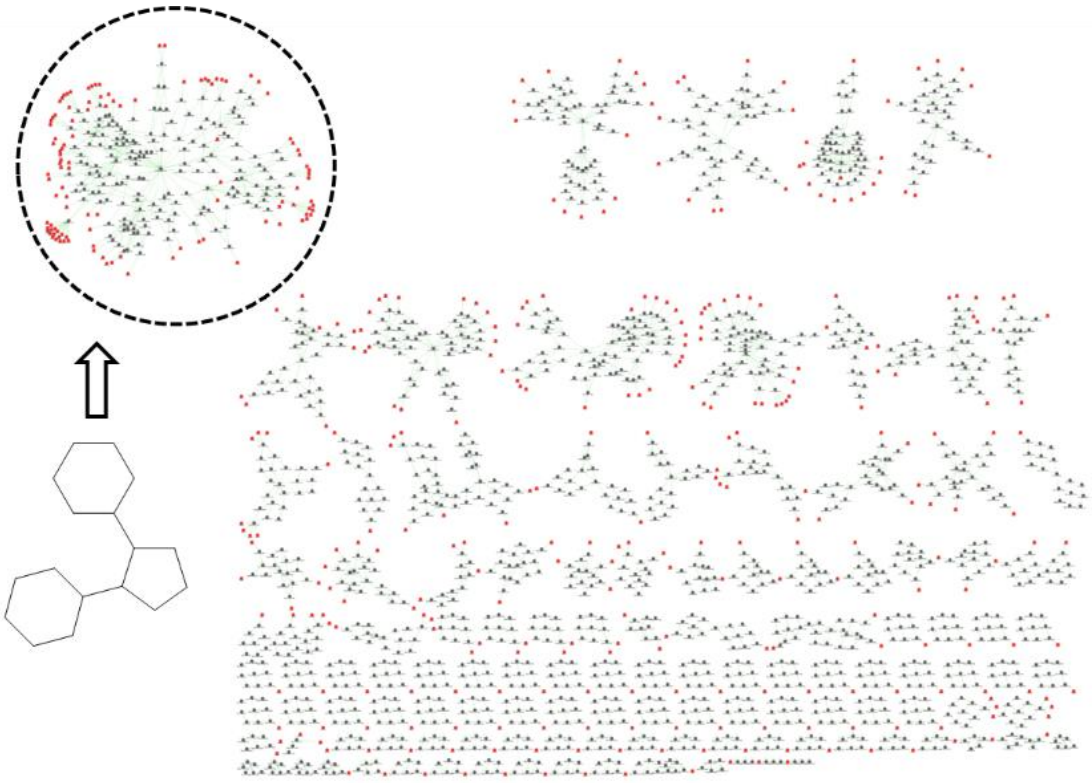


Figure 8. Cystoscape Plot of the clusters generated using the Molecular Assemblies metrics on 454 COX2 Inhibitors. Highlighted the most relevant cluster accounting for the "coxib" class of molecules

Chapter 6 reports an application of MA in the design of a diverse drug-like scaffold based library optimized for HTS campaigns. A well designed, sizeable and properly organized chemical library is a fundamental prerequisite for any HTS project. To build a collection of chemical compounds with high chemical diversity was the aim of the Italian Drug Discovery Network (IDDN) initiative. A structurally diverse collection of about 200,000 chemical molecules was designed and built taking into account practical aspects related to experimental HTS procedures. Algorithms and procedures were developed and implemented to address compound filtering, selection, clusterization and plating.

Chapter 7 collects concluding remarks and plans for the further development of the tools

Development and Validation Of In Silico Tools For Efficient Library Design and Data Analysis in High Throughput Screening Campaigns

Chapter 3: Ligand Generator A De Novo Tool For Structure Based Virtual Screening

3 LIGAND GENERATOR A DE NOVO TOOL FOR STRUCTURE BASED VIRTUAL SCREENING

3.1 INTRODUCTION

Computer Aided Drug Design and Development (CADDD) is an effective strategy for accelerating and economizing the drug discovery and development process as reported in Chapter 1. Because of the dramatic increase in the availability of biological macromolecule and small molecule information, the applicability of computational drug discovery has been extended and broadly applied to nearly every stage in the drug discovery and development workflow, including target identification and validation, lead discovery and optimization and preclinical tests[1].

Molecular de novo design aims to generate novel chemotypes endowed with a particular set of desired pharmacological properties. Indeed the properties have to be tuned specifically according to the biological target localization, the administration route, the metabolic profile[2]… The *de novo* design process is ideally driven by knowledge, knowledge is needed to define constraints to which the novel molecular structures being generated are supposed to obey. As extensively discussed by Schneider and Fechner[2], the search for novel chemotypes is a local and a multi-objective optimization process[3], which can lead to several solutions according to the input constraints and to the generation rules chosen by the chemist. On the other hand, finding a globally optimal solution is unworkable given the enormous size of the available chemical space[4]. Therefore, computer programs have been actively sought during the past two decades to help (medicinal) chemists to find the best local solutions in terms of pharmacological properties, chemical feasibility, innovation, and patentability. As a result, several molecular de novo programs, such as LUDI [5], LEGEND[6], LeapFrog[7], LigBuilder 2[8], SPROUT[9], HOOK[10], PRO-LIGANDS[11], DOGS[12] have been developed and thorough reviews of the de novo programs are available[2, 13]. Despite several successful applications having been reported,[14] many problems eventually prevented molecular de novo design approaches to become established tools in drug design, and, in fact, methods like virtual screening and molecular docking have received greater attention, either in terms of successful applications or widespread use. The most relevant issues associated with these early de novo design methods are: i. Producing chemical structures with poor drug-like properties and without any means of synthetic feasibility:

- Poor synthetic accessibility of the suggested ligand[15];
- low structural diversity[14];
- Low potential for parallel synthesis applications (a part from when combinatorial chemistry is directly addressed[16]);

Structures have the tendency to be bigger than necessary and too functionalized, because of the attempt of the algorithm to optimize all the possible interactions of the ligand with the binding site.

The issue of synthetic feasibility, in particular, is the most relevant one and its impact on the outcome of the de novo approach is highly dependent on the stage of the project and on the size of the library that has to be assembled. When the main goal is an HTS primary screening campaign, many docking as well as *de novo* design programs are able to handle combinatorial explosions.

Using a simplified procedure: the definition of a set of fragments (reagents) with predefined anchor points, and trivial linkage rules can lead to the generation of targeted libraries of arbitrary size which can in

principle be synthesized in parallel. In this context, the ability to control chemistry, in terms of reaction steps and reagent's availability, is crucial to moving the designed libraries from virtual to real.

Another common application of de novo design is lead optimization. In this case, the binding mode of the core structure of the lead has usually already been validated and the scope of the *de novo* approach is to optimize decoration towards increased affinity and/or improved ADME properties. The value of the class under study is therefore high, and the number of compounds to be generated is reasonably low. Thus, the selection of the fragments is driven by the optimization process and not by the accessibility of the pre-synthesized reagents. Even in this case, a simplified set of rules for the assembly of the fragments to the core scaffold is generally used: synthetic feasibility is not addressed by the software itself, but by the medical chemist who defines the anchor points and the set of functional groups that will define the chemical space to explore.

Generally used also for scaffold hopping, Ligbuilder[8] and FlexX[17] in combination with ReCore[18], are able to generate new scaffolds for the same pattern of substituents.

A relevant application is the design of ligand-based libraries without a predefined set of scaffolds. In this case it is imperative to extensively map the available chemical space by defining a wide set of chemical rules.

Two approaches try to address the synthetic feasibility problem: the first is to use a retro-synthetic analysis of the final ligand, and this is generally done by the implementation of bond breaking rules. LigBuilder2[8] is an example in which an internal database of building blocks, while it is also possible to use standalone programs, like for example SYLVIA[19], developed by Molecular Networks. These approaches try to find and identity the bonds that more likely can be created by a chemical reaction and then break them. Like the Retrosynthetic Combinatorial Analysis Procedure (RECAP)[20] allows the identification of building block fragments rich in biologically recognized elements and privileged motifs and structures, this approach generates virtual fragments without any embedded evaluation their of commercial availability.

The second approach, used by most of the available de novo design programs uses restricted sets of chemical rules that encode for the most common chemical reactions. Some of the most relevant reaction driven de novo software's with an extended set of chemical reactions are SYNOPSIS[21] and DOGS[12].

In this scenario, we decided to develop a de novo design program able to employ a user-defined set of true chemical reactions as connection rules and to use tangible commercial reagents as building blocks. Ligand Generator (LiGen), described in this chapter, is a suite of programs able to define flexible de novo design workflows by using both pharmacophores and molecular docking engines, linking the fragments in situ, and directly scoring the drug likeness of the generated molecules.

In a standard application, the modules work sequentially, from the generation of the input constraints (either structure-based, through active site identification, or ligand-based, through pharmacophore definition), to docking and de novo generation. Alternatively, the modules can be used as standalone services according to the user's need. Figure 9 illustrates the general LiGen workflow.

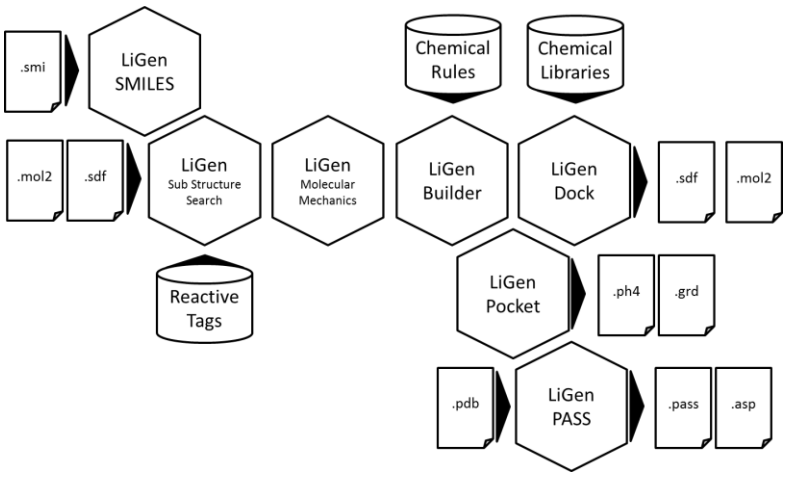


Figure 9.schematic representation of the LiGen modules and dataflow.

The structure and the function of individual modules are described in detail in the following paragraphs, along with a general description of the philosophy and of the functionalities and of the software. Identified issues and implemented solutions, during the 2014 research , will also be reported.

3.2 LiGen's Features

As mentioned above, the LiGen suite has been designed to be, at one and the same time, a modular *de novo* design application and a computational tool box. The LiGen modules can be used independently from each other and combined with other tools typically used in drug design research.

This is possible because of how the modules communicate each other by using a few standard data formats, with a few minor extensions for inter-module communication not breaking the original standard format. Moreover the modular structure of LiGen allows the implementation of different workflows using the basic building block tools, depending on the specific computation or analysis required.

LiGen is written in C++, with the use of STL (Standard Template Library) at high level, and its own algorithms (*functions*) and data structures (*classes*) at low level. In this way, the code is flexible, easily extensible and fast on many architectures. Performance portability to different hardware has been a main priority in LiGen design. Using STL at all levels (e.g. basic vector for interatomic distances) may result in a code not equally efficient on all hardware, or possibly not at all portable (think to GPU computing where CUDA or OpenCL kernels are required, and STL may not be fully supported). For these strategic reasons, code flexibility and portability to different hardware, Intel XEON PHI was selected as our platform of choice for HPC applications.

All LiGen modules share the same basic C++ classes representing main biochemical objects: Atoms, Residues, Molecules; or Mathematical objects: vector, grid, ring, etc. LiGen modules can be compiled in standalone executable or built together in a single executable, implementing a given workflow. The basic modules of LiGen are: LiGenPass, LiGenPocket, LiGenSmiles, LiGenMD, LiGenMinimizer, LiGenSSS, LiGenDock, LiGenScore, and LiGenBuilder.

3.2.1 Definition of the Input Constraints

The definition of the input constraints required to run a de novo design, given a biological target, is particularly relevant as it is directly connected to the quality evaluation of the candidate compounds which will eventually result from the run. In general, target constraints for the generated compounds are composed by a set of information related to ligand-receptor interactions. This information may equally derive from the known 3D structure of the ligand-receptor complex or from the structure of a set of ligands for that particular target. In the common case of receptor based approaches, input constraints are usually given in the form of shape constraints, derived from complementarity between ligand and binding pocket, and in the form of potential interaction points, usually classified as hydrogen bonds, electrostatic and hydrophobic interactions. The thus extracted constraints are used to score the candidate compounds.

The purpose of the scoring function is to delineate the correct poses from incorrect poses, or binders from inactive compounds in a reasonable computation time. However, scoring functions involve estimating, rather than calculating the binding affinity between the protein and ligand and through these functions, adopting various assumptions and simplifications[22]. Scoring functions can be divided into force-field-based, empirical and knowledge-based scoring functions[23].

Classical force-field-based scoring functions[24] assess the binding energy by calculating the sum of the non-bonded (electrostatics and van der Waals) interactions. The van der Waals terms are described by a Lennard-Jones potential function. Force-field-based scoring functions also present the problem of slow computational speed. Thus the cut-off distance is used to handle the non-bonded interactions. This also results in decreasing the accuracy of long-range effects involved in binding.

Extensions of force-field-based scoring functions consider the hydrogen bonds, solvations and entropy contributions. Software programs, such as DOCK[25], GOLD[26] and AutoDock[27], offer users such functions. In empirical scoring functions, binding energy decomposes into several energy components, such as hydrogen bond, ionic interaction, hydrophobic effect and binding entropy. Each component is multiplied by a coefficient and then summed up to give a final score. Coefficients are obtained from regression analysis fitted to a test set of ligand-protein complexes with known binding affinities.

Knowledge-based scoring functions[28] use statistical analysis of ligand-protein complex crystal structures to obtain the interatomic contact frequencies and/or distances between the ligand and protein. They are based on the assumption that the more favorable an interaction is, the greater the frequency of occurrence will be. These frequency distributions are further converted into pairwise atom-type potentials. The score is calculated by favoring preferred contacts and penalizing repulsive interactions between each atom in the ligand and protein within a given cutoff.

Consensus scoring[29] , the type implemented in LiGen, is a recent strategy that combines several different scores to assess the docking conformation. A pose of ligand or a potential binder could be accepted when it scores well under a number of different scoring schemes. Consensus scoring usually substantially improves enrichments in virtual screening, and improves the prediction of bound conformations and poses[29b].

LiGenPass is based upon the algorithm used by PASS (Putative Active Site with Spheres), software developed by Brady et al.[30] to identify cavities in the target proteins. As described in the Experimental Section, LiGenPass is a new implementation of the algorithm and it is used by the LiGen engine to define binding sites for the target protein. LiGenPass characterizes regions of buried volume in the target protein and identifies potential binding sites based upon the size, shape and burial extent of these volumes (Figure 10A).

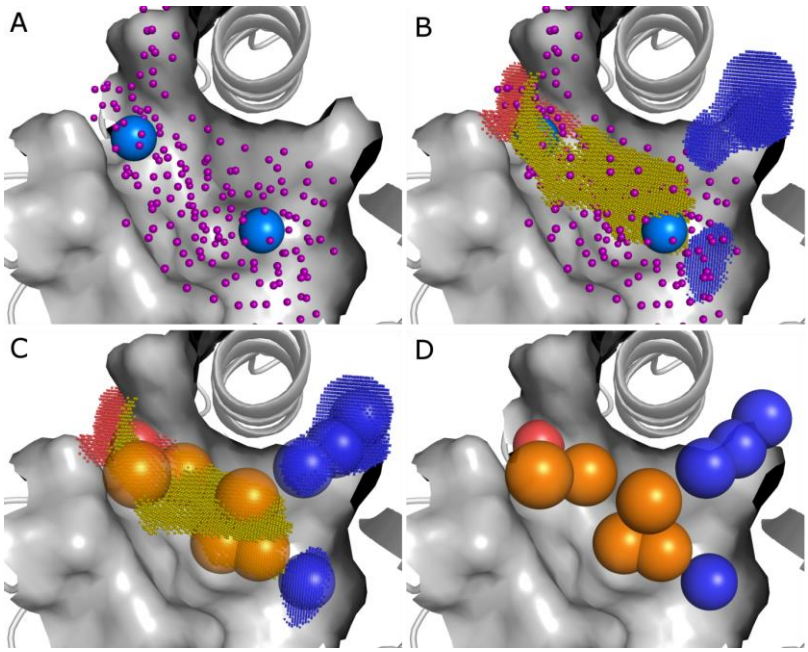


Figure 10. Identification of protein binding sites. A) LiGenPass maps the protein surface filling cavities with probes (violet); probes are clustered in ASP, representing the center of mass of the cavity. B) According with the tridimensional coordinates of ASP, the cavity surface is mapped in order to identify hydrophobic regions (yellow), H-bind donor (blue) and H-bond acceptor (red) regions (key sites). C) The key sites just identified are clustered into a pharmacophore model (in orange are represented the hydrophobic features, in blue the H-bond donor features, in red the H-bond acceptor feature. In figure D) is shown the pharmacophore model of the binding site. Protein PDB code: 1dfh. Figures are prepared with Pymol.[31]

Each cavity is defined by an Active Site Point (ASP) as the centre of mass of the cavity itself. The number of possible ASPs can be defined by the user as the threshold distance at which they are considered distinct units. Having identified a number of cavities through the corresponding ASPs, the user may want to characterize them in terms of key interaction points and suggest a pharmacophore which can subsequently be used for 3D database searching, docking procedure or *de novo* ligand computation.

This task is performed by the LiGenPocket module. In essence, LiGenPocket works by creating a regularly spaced grid around ASPs, if the target protein has no co-crystallized ligands, or by defining a sphere which covers the ligand and the protein atoms surrounding the ligand, and then creating a grid within the sphere, if there is a co-crystallized ligand.

The program then places a hydrogen atom as a probe on each grid point to check its accessibility. If the probe collides with the protein, that grid point will be labelled as 'not free'. A collision is counted when the interatomic distance is less than the sum of van deer Waals radii reduced by 0.5 Å. If the probe does not collides with the protein, that grid point will be labelled as 'free'. If a grid point is farther than 5 Å from any atom of the protein, it will be labelled as 'outside' rather than 'free'. The assembly of free grid points forms the body of the binding pocket in which each newly designed ligand will be built up.

As the next step, the program will derive key interaction sites within the binding pocket. Such information is necessary for the subsequent ligand construction or docking process. In the current implementation, the program uses three different types of probe atoms to screen the binding pocket: a positively charged sp3 nitrogen atom (ammonium cation), representing a hydrogen bond donor; a negatively charged sp2 oxygen atom (as in a carboxyl group), representing a hydrogen bond acceptor; a sp3 carbon atom (methane), representing a hydrophobic group. For each free grid point, the binding energies between the probes and the protein are calculated by using an in house developed scoring function. Each grid point will be labelled

as donor, acceptor, or hydrophobic according to the highest score achieved by one of the three probes on that particular point. The program will then filter all the grid points to derive the key interaction sites in a two-step process. In the first step, the program calculates the average score for all the grid points labelled as 'donor'. Then the program identifies grid points having a score lower than the average and labels them back as free. The same process is repeated for the 'acceptor' grid points and for the 'hydrophobic' grid points (Figure 10B). At this stage, LiGenPocket attempts a definition of pharmacophoric points. Each Surviving grid point, either donor, acceptor, or hydrophobic, is characterized for the number of 'neighbour' grid points. In this context, 'neighbours' are defined as grid points with the same definition (donor, acceptor, or hydrophobic), falling less than 2 Å from that particular point. The average of neighbours for each type of grid points is calculated, and those points having a number of neighbours lower than the average are labelled as free. After this step, only those grid points that 'aggregate' survive, and can be defined as the key interaction sites within the binding pocket. The geometric centre of each aggregation is calculated and finally defined as a pharmacophoric point (donor, acceptor, or hydrophobic) (Figure 10C).

This binding site-derived pharmacophore model can then be used for the subsequent docking of candidate compounds, or directly as a query structure to perform 3D database searching, which provides an additional way to find novel ligand molecules that fit to the target protein.

3.2.2 Docking Procedures

In a standard application of a de novo design approach, placement of fragments into the target protein according to a set of input constraints usually follows the characterization of the binding pocket(s). LiGen accomplishes this task by using the module LiGenDock, which can also be used as a standalone module to perform classical docking experiments. The main feature of LiGenDock is the use of the pharmacophore scheme generated by LiGenPocket as the driver for the docking procedure, including a non-enumerative flexible docking algorithm (Figure 11).

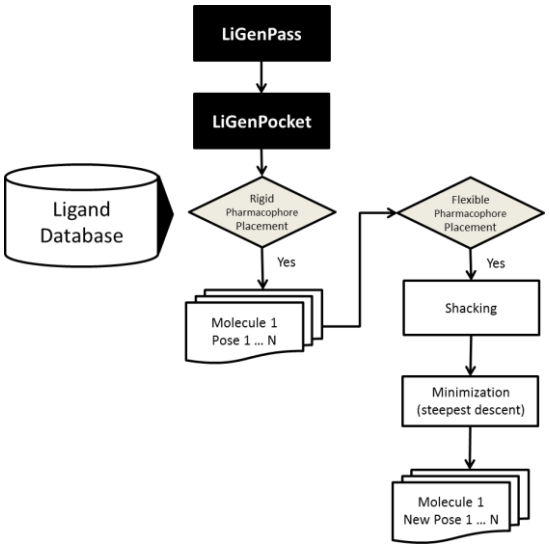


Figure 11. LiGenDock process workflow.

A through description of LiGenDock, along with its full validation and with the optimization of the docking parameters, is presented in a recent publication of our group[32].

3.2.2.1 Docking Algorithm

- Step 1. Aligns rigidly the center of mass of the ligand with the center of mass of the cavity (Figure 12B).

- Step 2. All possible combination pairs between the pharmacophoric points defined by LiGenPocket the interacting atoms on the ligand, in this step the ligand in considered as a rigid body (Figure 12C).
- Step 3. A conformer search is performed, allowing all torsional angles to rotate freely. The conformers with the best scores are retained for the successive step (Figure 12D).
- Step 4. loops over all ligand poses that match two pairs of pharmacophore feature-ligand atoms, trying to match a third pair by rotating the ligand around the axis connecting the two pairs already found (Figure 12D).
- Step 5. Exits the procedure by keeping the best poses of the ligand, which are then passed to an energy minimization, routine used to refine the position of the ligand inside the binding pocket (Figure 12E).

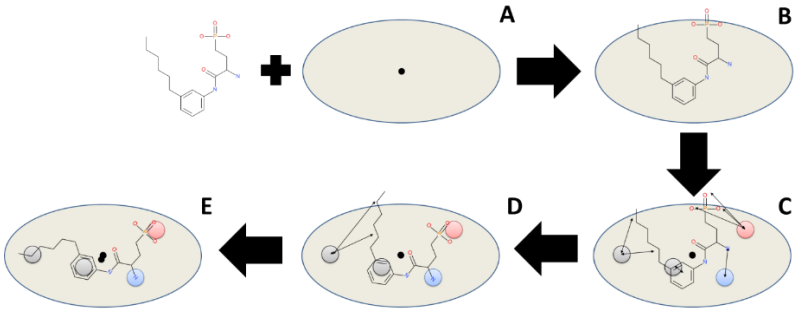


Figure 12. Cartoon of the workflow of the docking algorithm as originally implemented in LiGenDock. when the rigid placement of the ligand (Step 1 and 2) generates poses having a good overlap with the binding site and with the receptor's pharmacophore.

This algorithms has demonstrated good performance in a wide range of situations[32b], comparable with other docking programs such as Autodock, Glide and PLANTS[33].

In the routine use of the software in our lab, we found that the performance of LiGenDiock (using default parameters) with very flexible ligands in combination with particular binding pockets is very poor when trying to identify the correct pose of the ligand with respect to the experimentally determined one. In some cases the results greatly improve by tuning, the parameters of the docking engine do control the flexible part of the docking procedure (steps 3 and 4). The novel set of the parameters has the drawback of dramatically increasing the docking time (from 30 seconds per molecules to several minutes).

Therefore, we decided to investigate this set of "difficult" cases. After an extensive exploration of the docking parameters, we were able to isolate the problem. In Figure 13, we can see a schematic representation of the occurring issue. When the 3D conformation of the ligand differs too much from the shape of the cavity of the binding site the first rigid placement of the ligand (step 1 and 2) produces starting poses presenting too many clashes with the target protein and with a suboptimal alignment of the ligand with the receptor pharmacophore (Figure 13B). In this scenario, the generation of a good pose is devoted to the flexible part of the docking algorithm (Figure 13D) (step 3 and 4) and the optimal set of parameters is dependent on the specific ligand/site complex, this specificity prevents the implementation of a set of parameters able to correctly address this particular condition. In addition the step 3 and 4 of the docking process: the flexible pharmacophore refinement of the pose, are meant to find local minima, in this circumstance the algorithm is forced to find a global minima. The result is that the docking time increases 3 to 5 fold, with an average docking time shifted from 30 seconds to 2 minutes per molecule, this docking time is too high to use the software for HPC applications.

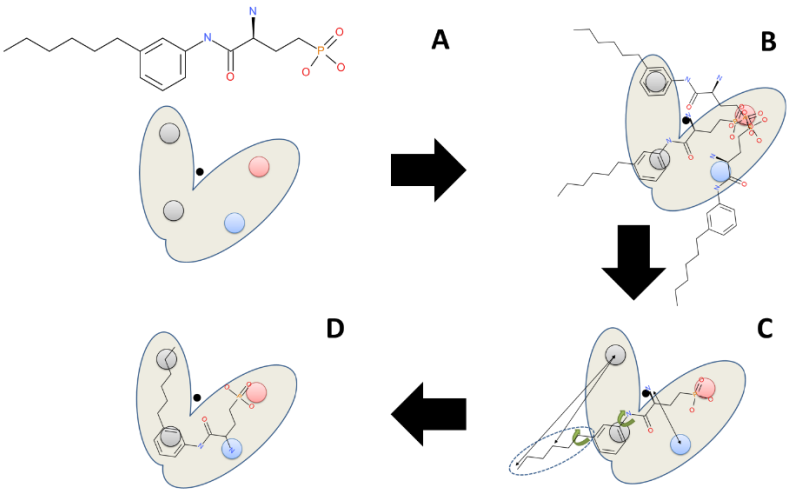


Figure 13. Cartoon of the workflow of the docking algorithm as originally implemented in LiGenDock, when the rigid placement of the ligand (Step 1 and 2) generates poses having a poor overlap with the binding site and with the receptor's pharmacophore.

We decided to address this limitation of the docking algorithm by redesigning from scratch the rigid placement of the ligand (steps 1 and 2). We found an interesting opportunity in the shape of complementarity methods.

Geometric matching/shape complementarity methods describe the protein and ligand as a set of features that make them dockable[34]. These features may include molecular surface/complementary surface descriptors like LigandFit[35]. In this case, the receptor's molecular surface is described in terms of its solvent - accessible surface area and the ligand's molecular surface is described in terms of its matching surface description[36]. The complementarity between the two surfaces amounts to the shape matching description that may help find the complementary pose of docking the target and the ligand molecules. Another approach is to use a Fourier shape descriptor technique[37]. Whereas the shape complementarity based approaches are typically fast and robust, they cannot usually model the movements or dynamic changes in the ligand/ protein conformations accurately, although recent developments allow these methods to investigate ligand flexibility.

Shape complementarity methods can quickly scan through several thousand ligands in a matter of seconds and actually figure out whether they can bind at the protein's active site, and are usually scalable to even protein – protein interactions[38].

By what means shape complementarity methods are good in the prediction of a bioactive conformation, has been recently assessed by Renxiao Wang group, Li et al[39] tested a panel of 20 scoring functions. They introduced a naïve scoring function as a reference, the buried solvent accessible surface area of the ligand molecule upon binding (ΔSAS). ΔSAS is an estimation of the size of the protein− ligand binding interface. They observed that ΔSAS outperformed most scoring functions in the scoring test[39]. This embarrassing fact indicates that current scoring functions have not advanced too far from very simple accounts of protein−ligand interactions. It also suggests that the major problem in computing ligand binding affinities does not lie in the nonpolar aspect. Nonpolar interactions are conventionally modeled by the non-discriminative contacting surface area, and many scoring functions actually have such an account explicitly or implicitly. Instead, polar interactions, as well as the desolvation effect associated with them, remain as the real challenge. Besides, its performance in the docking of the directory of useful decoy (DUD) database[40] is virtually the worst among all scoring functions under evaluation. ΔSAS is a descriptor of the nonpolar factors in protein−ligand interactions, which are non-discriminative in nature. Therefore, it is not

really capable of differentiating the decoys from the true binding pose since they are in fact the same ligand molecule. To achieve a good docking power, a scoring function needs necessary consideration of specific, directional polar interactions, such as hydrogen bonds.

For our scope, the reimplementation of the algorithm for the identification of the initial pool of starting poses, the limitation highlighted by Li et al[39] is not relevant because the shape complementarity method is only the first step of LiGenDock. The identified poses are then flexibly aligned to the receptor pharmacophore (Step 3 and 4) and finally the interactions occurring with the receptor are evaluated by a consensus scoring function LiGenScore. Therefore, for us the relevant information is that this method is able to identify bioactive conformations.

Two new algorithm based on spatial geometry of the pharmacophore and topology of the ligand have been introduced to address two problems:

- the independency of the docking result from the initial configuration of the ligand, and the shape matching bedween ligand and volume available inside the protein pocket.
- The first algorithm is a minimization algorithm that perform an unfold of the ligand using all rotational degrees of freedom, with the objective of maximizing the sum of all distance between ligand atoms.
- Moreover there is the option to relax the atomic bonds and angles using mff94 force field data.

The second algorithm uses the probe atoms that have been used to define the pochet as the definition of the target volume to be maximally occupied by the ligand. The procedure is iterative and start computing the center of mass and moments of inertia of the probe atoms (using a fake atomic mass equal for all probes), then the algorithm do the same for the unfolded ligand (using the same fake atomic mass as for the probes, and regardless of the atomic type), after that the algorithm proceed aligning the center of mass of the ligands with the probes and then rotate the ligands so that the inertia axis of the ligand are aligned with the inertia axis of the probes.

Finally a procedure similar to the algorithm used to unfold the ligand is used to optimize the geometry of the ligand to match the magnitude of the inertia axis of the probe set. The optimal result is a ligand whose shape match those of the probe. Obviously this is an ideal situation, in reality, since the space of configuration has a lot of local minima, the procedure; most of the time will find a sub-optimal configuration with the solution in one of this local minima. The solution could be eventually improved using a simulated annealing procedure (or some other technique to find global minima), but the procedure risk then to become costly in term of compute time. A local minima, in this specific contest, is enough most of the time for our goal

3.2.3 Reproducibility of the docking results

One of the most relevant strength of the docking engine should be the reproducibility of the docking poses. The implemented algorithm in LiGenDock is completely deterministic. To test the stability of the engine we have rerun LiGenDock (original implementation of the algorithm) with the same parameters and the same 3D structure (((3R)-3-amino-4-[(3-hexylphenyl)amino]-4- oxobutyl)phosphonic acid as bound to S1PR1 (PDB ID: 3V2Y)) of the same ligand 455 times. Results are reported in Figure 14, the software generates exactly the same pose in all the runs.

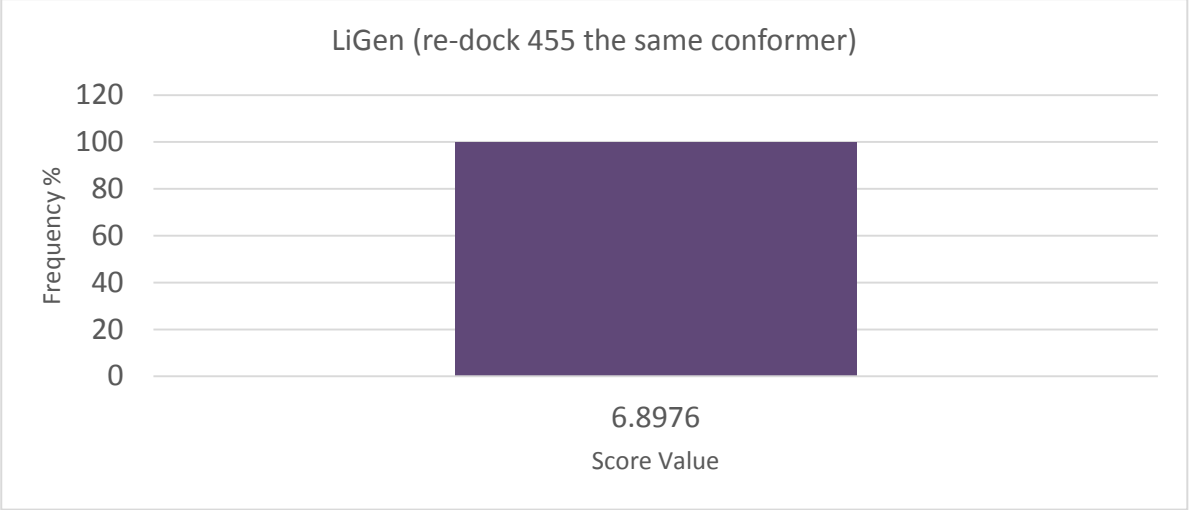


Figure 14. Score values distribution plot generated by LiGen in the re-dock of 455 the same conformer test

To compare these results with other docking programs: we report in Figure 15 the results of VINA software and in Figure 16 the Autodock results.

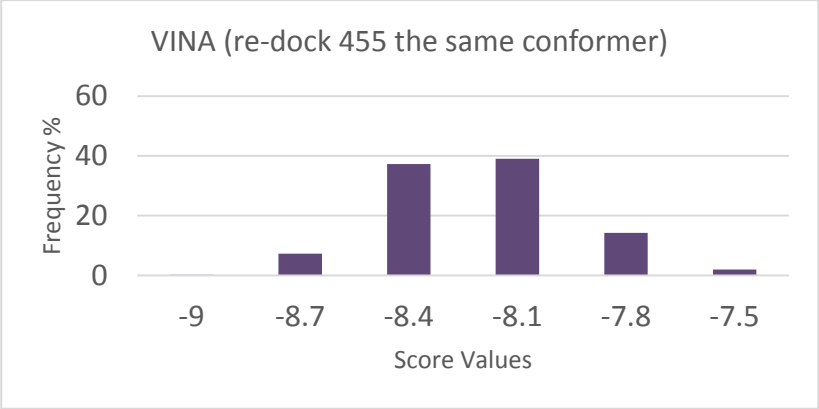


Figure 15. Score values distribution plot generated by VINA in the re-dock of 455 the same conformer test

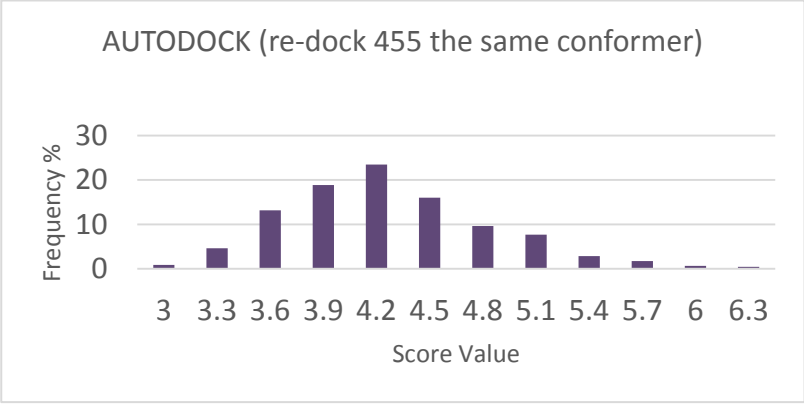


Figure 16. Score values distribution plot generated by Autodock in the re-dock of 455 the same conformer test

The stochastic docking engine of VINA generate an ensemble of solutions with a standard deviation of 1.25 units of score values and Autodock's score values spread in a range of 3.3 score units.

The stability of LiGen re-dock versus the other docking software is for obvious reasons a remarkable feature.

To test the dependency of the software from the starting conformation, we selected a very flexible ligand with nine rotatable bonds.

We generated for the (((3R)-3-amino-4-[(3-hexylphenyl)amino]-4- oxobutyl)phosphonic acid as bound to S1PR1 (PDB ID: 3V2Y) 455 diverse conformers. Then we docked the ligand conformations in the reference target with LiGen, original implementation of the algorithm (Figure 17), VINA (Figure 18) and Autodock (Figure 19).

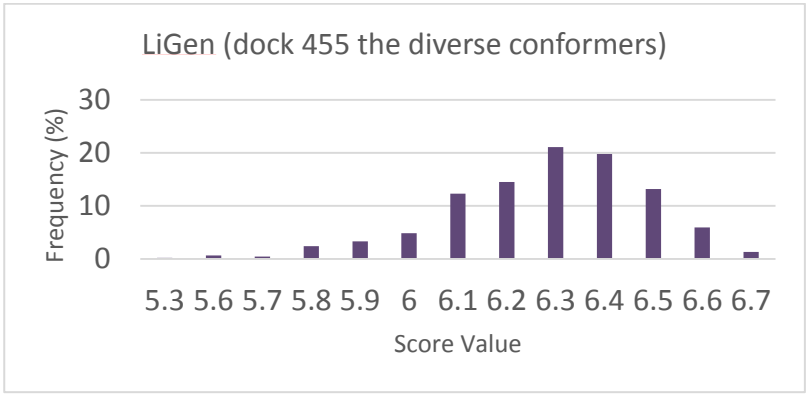


Figure 17. Score values distribution plot generated by LiGen in the dock of 455 the diverse conformer test

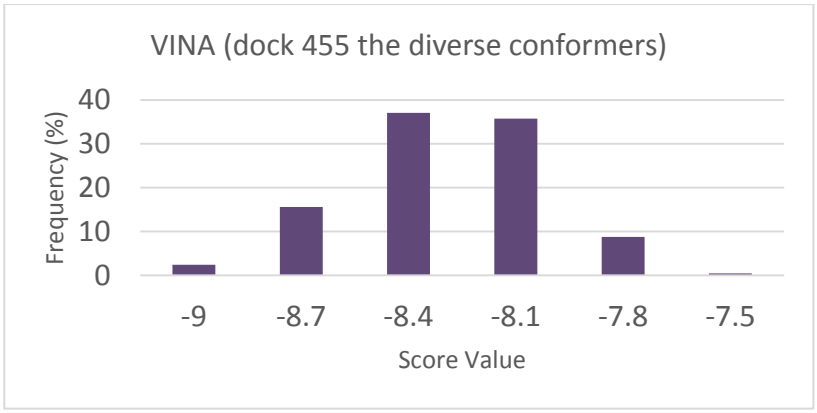


Figure 18. Score values distribution plot generated by VINA in the dock of 455 the diverse conformer test

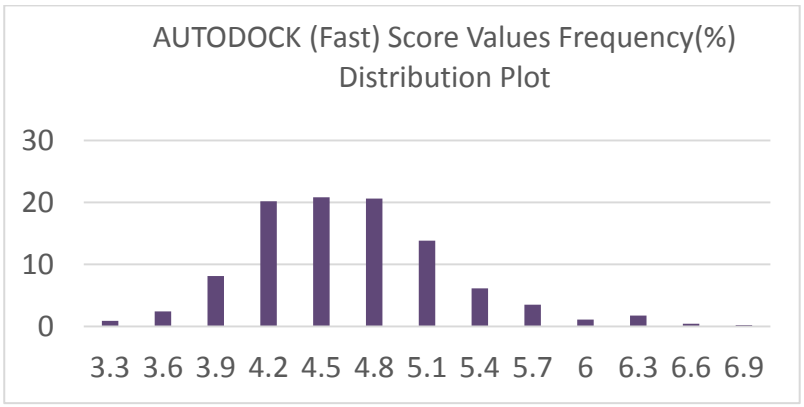


Figure 19. Score values distribution plot generated by VINA in the dock of 455 the diverse conformer test

The effect of the starting conformation on the range of values in the scoring function is 1.4 with LiGen, 1.5 with VINA and 3.6 with AutoDock. VINA in the two stability tests (re-dock of the same conformer 455 times and dock of 455 most diverse conformers), produces almost the same score frequency distribution plot (Figure 20). It means that, the bias introduced by the stochastic search has the same effect and magnitude on the results as a careful sampling of the conformational space of the ligand.

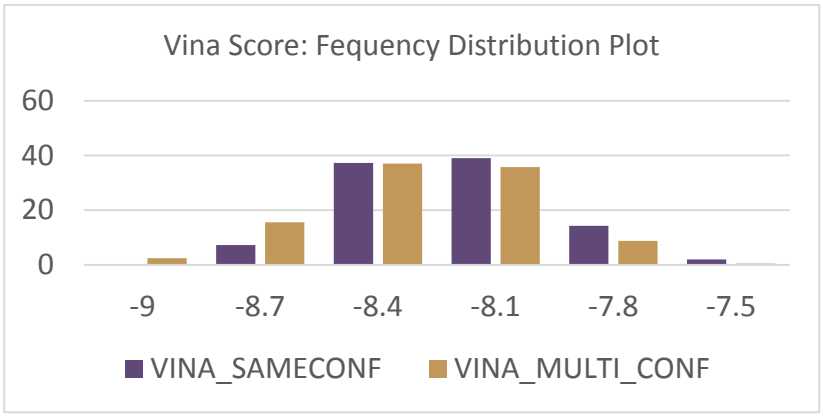


Figure 20. Score values distribution plot generated by VINA for the two stability tests

Even if LiGen has the lower oscillation in the results, it is nevertheless unacceptable for our stability criteria.

In a virtual screening campaign, where the n-top percent is selected for the actual experimental validation, the initial conformation of the ligand introduces a strong bias to the results. This bias in LiGen could be reduced if an ensemble of conformation is evaluated instead of a single conformation, (for the other software evaluated, the fluctuation of the results is intrinsic in the stochastic nature of the docking algorithm and cannot be avoided). LiGen is meant for massive virtual screening campaigns and by design, it is implemented to avoid conformational sampling of the standalone ligand.

As mentioned in the previous paragraph, we decided to remove this bias by the implementation of a novel algorithm based on shape complementarity method. Preliminary results on the reference ligand receptor complex show that the conformational bias can be reduced at 0.6 score units the variability range of the results; generate better results with a window of values between 6.8 and 6.2.

3.2.4 Improving Pose/Compound Selection

Due to the poor performance of current scoring functions in estimating binding affinity and hence in ranking docked ligands, it is recognized that compound selection based on calculated scores is not sufficient and visual inspection is often necessary. However, a practical concern arises if one needs to manually inspect thousands of docking poses[41]. Therefore, huge efforts have been devoted to automating this procedure based on the indications gained from ligand–target interactions. The molecular interaction fingerprints (IFPs), which are simple bit strings that encode 3D information about ligand– target interactions into 1D binary vector, have been extended by Marcou and Rognan as a post-docking filter to prioritize the most relevant poses of low molecular weight fragments[42]. In their study, IFPs were evaluated with four popular docking tools (FlexX, Glide, Gold, and Surflex) for extracting the scaffolds of true CDK2 inhibitors. They observed that scoring by the Tanimoto similarity of IFPs to a given reference was statistically superior to conventional scoring functions in placing the low molecular weight fragment in the CDK2 binding site. Based on the assumption that active compounds should have specific contacts with their target to display activity and also to tackle the inefficiency of traditional clustering of docking poses, Bouvier et al. have proposed the Automatic analysis of Poses using a Self-Organizing Map (AuPosSOM)

method for pose ranking with careful analysis of interatomic contacts between the docked ligand and the target[43]. They have demonstrated that it is possible to differentiate active compounds from inactive ones using only mean protein contacts' footprints calculated from the multiple conformations given by docking software.

In this scenario, we decided to embed directly in docking procedure the use of the inverse pharmacophore (the pharmacophore generated by the receptor interaction maps), as described in the paragraph Docking Procedures (3.2.2). Matching the receptor's pharmacophore assures that in all the selected poses the chemical features are complementary to the key interaction site in the binding pocket.

During the third year of my PhD, we further extended and exploited the use of the pharmacophoric features mapping.

This descriptor can be used in data post-processing to select the subset of poses / ligands engaging specific interactions with the binding site. Pharmacophoric features mapping can be used as knowledge based on a hard filter of compounds to not establish required interactions with the receptor for example with enzymes. Alternatively, it can be used to create a model able to further enrich the number of active compounds in virtual screening campaigns selecting compounds able to interact with the same pharmacophoric points of the reference active ligands. At this stage, we are evaluating recursive-partitioning methods to generate discriminative models from the inverse pharmacophoric features mapping (Figure 21).

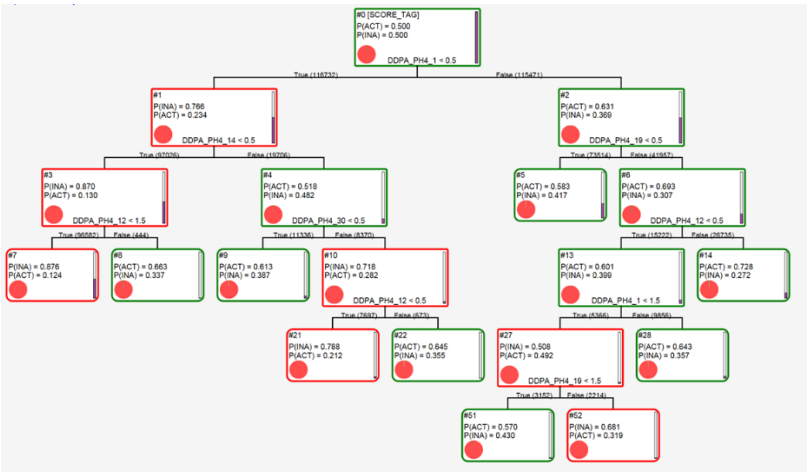


Figure 21. Recursive partitioning model based on the inverse pharmacophore features mapping

The results, on the effect of the combined model scoring function /recursive partitioning, we have obtained so far are too preliminary to be presented hire.

3.2.5 Structure Generation and Synthetic Accessibility

Several strategies can be adopted to sample and generate structures by de novo approaches. They are usually referred to as linking[5, 8], growing[44], lattice-based sampling, random structure mutation, transitions driven by molecular dynamics simulations, graph-based sampling[2], the former two having received particular attention. The linking approach starts with fragments positioned or pre-docked at different interaction sites of the binding pocket. Then, the fragments are linked together to give a molecule able to satisfy to the greatest extent all the interaction points. Also the growing approach starts with a positioned fragment which is grown to maximize favourable interactions with the key interaction sites or with receptor regions in between key interaction sites.

The basic building blocks for the assembly of candidate structures can be either single atoms or fragments. Atom-based approaches are superior to fragment-based methods in terms of the structural variety that can be generated, but this increase in potential solutions makes it harder to find feasible and druggable candidate compounds among the ones that have been identified. LiGen implements the structure building in a way that effectively ensures the synthetic feasibility of the candidate compounds. The building model of LiGen, called LiGenBuilder, is at the heart of LiGen and is the main engine implementing our de-novo algorithm. In particular, LiGenBuilder is combined with the docking (LiGenDock) and the force field engine (LiGenMD) modules to cooperate in order to perform in situ de novo growth of novel ligands. The building algorithm implemented in LiGenBuilder is described in detail in the next paragraph.

3.2.5.1 De Novo Design

In Figure 22.The overall process of ligand growth is presented.

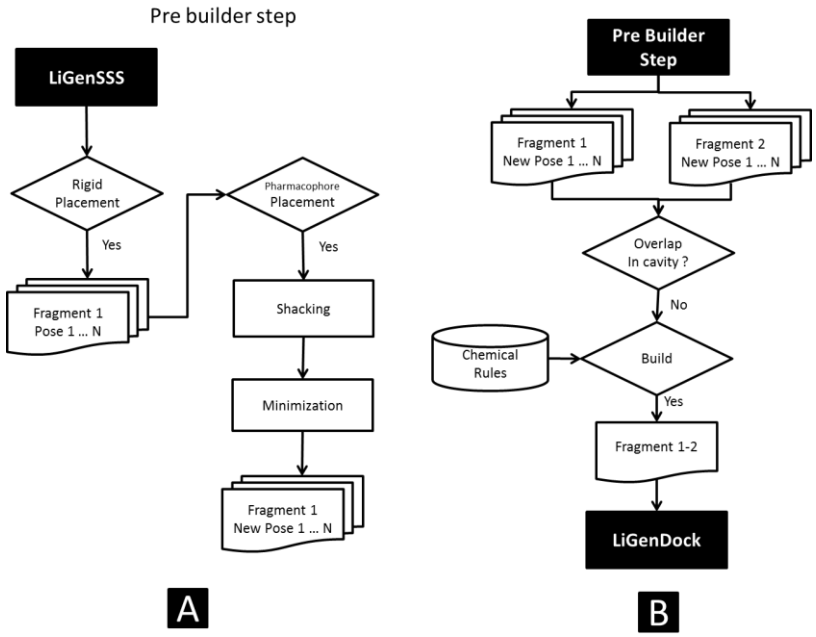


Figure 22. De novo design process of LiGen. Black boxes represents modules before LiGenBuilder

The process starts with a pre-building step (Figure 22.A) where a set of fragments is docked into the target active site and a number of poses that match given criteria are retained into a set which will be used as a basis for the successive growth process. In detail, seed building blocks are assigned with force field parameters (MMFF94[45]) and minimized (geometry optimization using MMFF94 force field). The building block atoms are also tagged for all possible reactions defined in a *chemical rules database*. Seed building blocks are then passed through three sub-phases. The first one is rigid docking, where the fragments are placed inside the pocket using the scoring function and the best poses kept in the seed set of poses; then the torsional angle of the poses are allowed to move so that the poses can be further optimized to lower the binding energy with the target protein; finally the complexes between poses and the target protein are minimized to find a local energy minimum. A pose is kept inside the set of best poses if the difference between internal energy before and after the minimization inside the pocket is less than a given threshold (defined by the user).

After this starting set has been completed, the process continues with a sequence of building steps (Figure 5B) where the poses contained in the starting set are grown by using building blocks selected randomly

(according to a probability function that depend on the chemical rules) among all the fragments that can react with the pose.

The new fragments thus formed are then docked again inside the protein pocket and the growing cycle continues. The process ends when either all the possible combinations of fragments have been explored (very rare event) or the molecular weight of the candidate ligand (pose) is larger than a user-defined value (Figure 23).

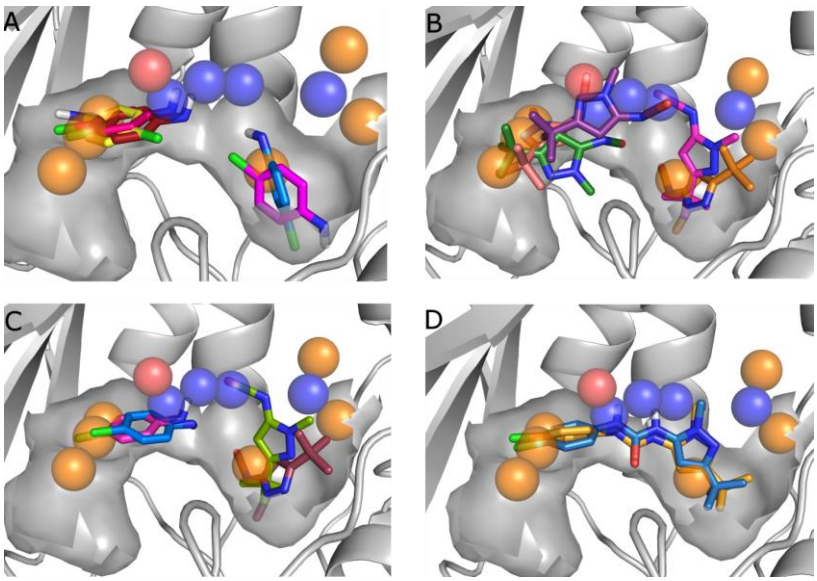


Figure 23. Example of ligand growing process. A) and B) placement of the fragments inside the binding site in order to match the features of the pharmacophore; C) different fragments shown together in the binding site, matching different pharmacophore features; D) the final generated ligand is minimized inside the binding site in order to catch the most favorable interactions (in orange is represented the generated ligand, in blue the co-crystallize ligand pose). PDBcode of the example complex: 1kv1. Figures are prepared with Pymol[31].

3.2.5.2 Chemical Rule Database

To perform the linking of two fragments or building blocks, a database of the most frequently established chemical reactions in parallel and combinatorial chemistry is used. The database has been encoded using the MDL mol format tagging the atoms with labels.

Two kinds of labels have been implemented: the part of the molecule that will be not present in the product is tagged with "<Clip>" and the reactive point for the reaction is labelled with an atom environment description tag. For example a reactive chlorine on a phenyl ring is tagged as "<Ph.Cl1>". Other examples of tagged structures are presented in Figure 24

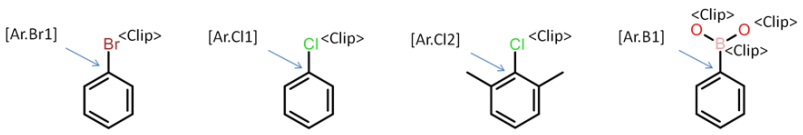


Figure 24. Examples of tagged chemical structures.

The key step of the de novo design process is now the transfer of the set of reactive points to the set of fragments or building blocks. Unlike other software like LigBuilder[8], LiGen does not need a pre-tagged database of fragments. On the contrary, an arbitrary database of fragments can be used without manipulation in the de novo process. This feature is provided by the LiGenSSS module (Figure 25).

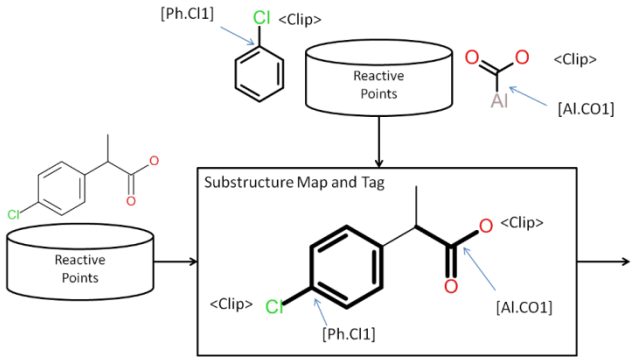


Figure 25. Example of how LiGenSSS works. It transfers tags from the reactive point database to the fragment database used in the de novo design process.

LiGenSSS carries out a substructure search in the database of fragments and identifies the reactive part of the building block. Once a match between a fragment substructure and a reactive point is found, LiGenSSS transfers the tag(s) from the reactive point database to the matched structure in the fragment database. As a result, a database of tagged reagents (building blocks) is available.

The assembly of reactants is carried out, in a stepwise process, by LiGenBuilder. LiGenBuilder reads a text file where a simple linear notation for all the possible defined chemical reactions is encoded. Figure 26 shows the part of this reaction file for a Suzuki coupling reaction.

```
BEGIN_REACTIONS
# The 1st  Column:      Atom-Tag of Fragment A
# The 2nd  Column:      Atom-Tag of Fragment B
# The 3rd  Column:      Bond Type Generated
# The 4th  Column:      Functional Group ID
# The 5th  Column:      Distribution of the functional group of each reaction
# The 6th Column:       Comments

# Suzuki Coupling

@           @       C-C Bond        1       0.50    ...
Ar.Br1      Ar.B    C-C Bond        1       0.60    ...
Ar.Cl1      Ar.B    C-C Bond        1       0.30    ...
Ar.Cl2      Ar.B    C-C Bond        1       0.10    ...
```

Figure 26. Example of a reaction file (a Suzuki coupling reaction).

The user can edit the file to encode pairs of reacting groups. It is also possible to control the distribution of the chemical reactions in the library. For instance, column five of the reaction file controls the probability that a given reaction will be cast by the LiGenBuilder grow engine. According to the example given in Figure 26, a Suzuki coupling has a 50% probability of being selected. Among all the possible Suzuki couplings, the aryl-Bromine derivative has a 60% chance of being selected versus the other two reactants. On the other hand, the sterically hindered chlorine derivatives [Ar.Cl2] will be used only if reagents with a higher reactivity are not available.

Useing defined casting rules allows guidance of the chemistry of the library, for example limiting the number of synthetic routes or focusing first on compounds having the higher chance to be successfully synthetized.

DISCUSSION

Despite their theoretical very high relevance, de novo design methods did not become established tools in drug design because of the presence of several drawbacks that limited their application to real world cases. LiGen, the suite of programs described in this paper, aims at addressing most of the issues associated with available de novo design approaches. Key features of LiGen are: i) pharmacophoric constraints are implemented in the pose identification. This allows the obtaining of only the poses endowed with the correct pattern of interactions with the binding pocket, and therefore significantly reduces the time for filtering them out once generated, as it is done by most of the docking programs and time is needed for visual inspection. (ii) Flexible docking without conformer enumeration. Even for molecules with a few rotatable bonds, the coverage of the conformational space accomplished by enumerative methods, is only partial. Thus, docking methods based on conformational sampling are biased by the degree of coverage of the conformational space. Flexible alignment of the molecule on the pharmacophore points eliminates this bias and produces the bioactive conformation of the molecule, as long as the interaction points are correctly defined. If energetically allowed, this conformation is retained. If not, the molecule is discarded as unable to match the possible pharmacophoric mappings. This minimizes the number of molecular conformations which are evaluated for each ligand. A detailed description and optimization of the docking routines is presented in an accompanying paper. (iii) Accurate and flexible reactant mapping. Although many excellent de novo design softwares effectively handle chemical rules (for example DOGS[12]), LiGen stands out by allowing the user to prioritize the chemical reactions through the definition of a probability of a reaction class (in the example, Suzuki coupling has a 50% of chance of being selected among all other reactions) and the definition of a probability of a reactive group meaning that if a building block is present such as bromine and chloride derivatives the most reactive one has more chance of being selected by the system to form the final molecule. This fine tuning allows the smallest number of chemical reactions to synthesize the *de novo* library with the selection of the most promising building blocks in terms of success of synthesis and yields.

(iv) Reactant tagging through substructure searching. In this way, the library of building blocks, fragments, or reagents, does not need to be manually tagged. Having defined the reactive chemical points, a substructure search routine automatically transfers the tag(s) to the fragment library. The identification of novel reactive points of the definition of subgroups is straightforward.

(v) Modular architecture. Although LiGen was intended to be a suite of programs aimed at de novo design, it can also be used in a modular fashion, thus allowing each module to be used as a standalone tool, or to combine different modules to assemble task-specific workflows. For example it is possible to define a workflow just to enumerate a *de novo* library according to defined chemical rules and used again the LiGenSSS to filter out unwanted combination of building blocks Figure 27.

Interestingly, the LiGen source code has been written in such a way that a given workflow can be composed either by a pipe of different executable or can be compiled in a single embedded binary executable Figure 27.

In conclusion, we presented a new method for *de novo* design based on pharmacophore-driven docking and accurate chemical reaction mapping and implemented to be suitable for HPC applications and portable to many hardware architectures.

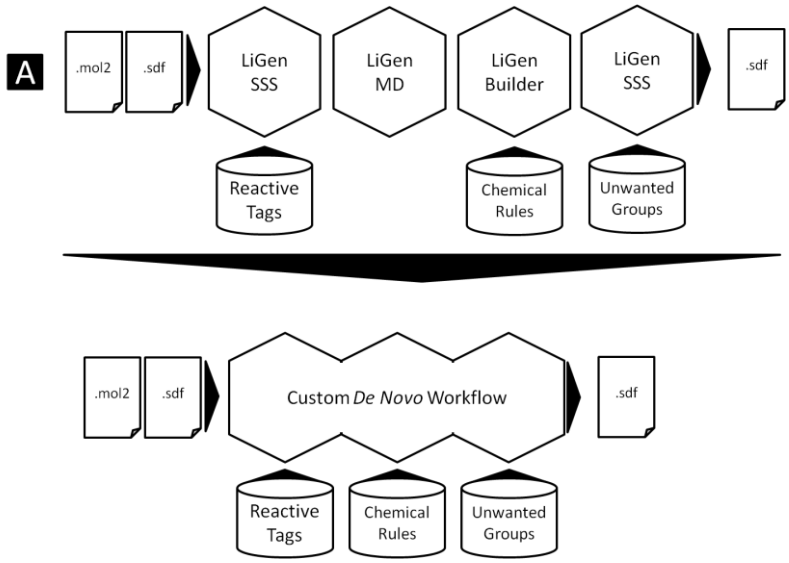


Figure 27. Example of a user-defined LiGen workflow. In this example the workflow was composed to enumerate a de novo *library according to user-defined chemical rules and to exclude from this library unwanted combination of building blocks (i.e. possible new molecules) running again, after LiGenBuilder, the LiGenSSS module.*

3.3 REFERENCES

1. Ou-Yang, S. S.; Lu, J. Y.; Kong, X. Q.; Liang, Z. J.; Luo, C.; Jiang, H., Computational drug discovery. *Acta pharmacologica Sinica* **2012,** *33* (9), 1131-40.
2. Schneider, G.; Fechner, U., Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov* **2005,** *4* (8), 649-63.
3. (a) Dey, F.; Caflisch, A., Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J Chem Inf Model* **2008,** *48* (3), 679-90; (b) Nicolaou, C. A.; Apostolakis, J.; Pattichis, C. S., De novo drug design using multiobjective evolutionary graphs. *J Chem Inf Model* **2009,** *49* (2), 295-307; (c) Nicolaou, C. A.; Kannas, C.; Loizidou, E., Multi-objective optimization methods in de novo drug design. *Mini Rev Med Chem* **2012,** *12* (10), 979-87.
4. Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L., Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J Chem Inf Model* **2012,** *52* (11), 2864-75.
5. 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* (1), 61-78.
6. Honma, T.; Hayashi, K.; Aoyama, T.; Hashimoto, N.; Machida, T.; Fukasawa, K.; Iwama, T.; Ikeura, C.; Ikuta, M.; Suzuki-Takahashi, I.; Iwasawa, Y.; Hayama, T.; Nishimura, S.; Morishima, H., Structure-based generation of a new class of potent Cdk4 inhibitors: new de novo design strategy and library design. *J Med Chem* **2001,** *44* (26), 4615-27.
7. Tan, J. J.; Zhang, B.; Cong, X. J.; Yang, L. F.; Liu, B.; Kong, R.; Kui, Z. Y.; Wang, C. X.; Hu, L. M., Computer-aided design, synthesis, and biological activity evaluation of potent fusion inhibitors targeting HIV-1 gp41. *Med Chem* **2011,** *7* (4), 309-16.
8. Yuan, Y.; Pei, J.; Lai, L., LigBuilder 2: A Practical de Novo Drug Design Approach. *J Chem Inf Model* **2011**.
9. (a) Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P., SPROUT: recent developments in the de novo design of molecules. *J Chem Inf Comput Sci* **1994,** *34* (1), 207-17; (b) Sova, M.; Cadez, G.; Turk, S.; Majce, V.; Polanc, S.; Batson, S.; Lloyd, A. J.; Roper, D. I.; Fishwick, C. W.; Gobec, S.,

Design and synthesis of new hydroxyethylamines as inhibitors of D-alanyl-D-lactate ligase (VanA) and D-alanyl-D-alanine ligase (DdlB). *Bioorg Med Chem Lett* **2009,** *19* (5), 1376-9.
10. Eisen, M. B.; Wiley, D. C.; Karplus, M.; Hubbard, R. E., HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins* **1994,** *19* (3), 199-221.
11. (a) Westhead, D. R.; Clark, D. E.; Frenkel, D.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B., PRO-LIGAND: an approach to de novo molecular design. 3. A genetic algorithm for structure refinement. *J Comput Aided Mol Des* **1995,** *9* (2), 139-48; (b) Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R., PRO-LIGAND: an approach to de novo molecular design. 1. Application to the design of organic molecules. *J Comput Aided Mol Des* **1995,** *9* (1), 13-32; (c) Waszkowycz, B.; Clark, D. E.; Frenkel, D.; Li, J.; Murray, C. W.; Robson, B.; Westhead, D. R., PRO_LIGAND: an approach to de novo molecular design. 2. Design of novel molecules from molecular field analysis (MFA) models and pharmacophores. *J Med Chem* **1994,** *37* (23), 3994-4002.
12. Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G., DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol* **2012,** *8* (2), e1002380.
13. (a) Diana, R., Computer-Aided Molecular Design. In *Handbook of Chemoinformatics Algorithms*, Chapman and Hall/CRC: 2010; pp 295-315; (b) Ji, H., Fragment-Based Drug Design: Considerations for Good ADME Properties. In *ADMET for Medicinal Chemists*, John Wiley & Sons, Inc.: 2011; pp 417-485; (c) Proschak, E.; Tanrikulu, Y.; Schneider, G., Chapter 7 Fragment-based De Novo Design of Drug-like Molecules. In *Chemoinformatics Approaches to Virtual Screening*, The Royal Society of Chemistry: 2008; Vol. 0, pp 217-239.
14. Kutchukian, P. S.; Shakhnovich, E. I., De novo design: balancing novelty and confined chemical space. *Expert Opin Drug Discov* **2010,** *5* (8), 789-812.
15. Ertl, P.; Schuffenhauer, A., Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform* **2009,** *1* (1), 8.
16. Jia, Y.; Chiu, T. L.; Amin, E. A.; Polunovsky, V.; Bitterman, P. B.; Wagner, C. R., Design, synthesis and evaluation of analogs of initiation factor 4E (eIF4E) cap-binding antagonist Bn7-GMP. *Eur J Med Chem* **2010,** *45* (4), 1304-13.
17. (a) Stiefl, N.; Zaliani, A., A knowledge-based weighting approach to ligand-based virtual screening. *J Chem Inf Model* **2006,** *46* (2), 587-96; (b) Gastreich, M.; Lilienthal, M.; Briem, H.; Claussen, H., Ultrafast de novo docking combining pharmacophores and combinatorics. *J Comput Aided Mol Des* **2006,** *20* (12), 717-34.
18. Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M., Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J Chem Inf Model* **2007,** *47* (2), 390-9.
19. Zaliani, A.; Boda, K.; Seidel, T.; Herwig, A.; Schwab, C. H.; Gasteiger, J.; Claussen, H.; Lemmen, C.; Degen, J.; Parn, J.; Rarey, M., Second-generation de novo design: a view from a medicinal chemist perspective. *J Comput Aided Mol Des* **2009,** *23* (8), 593-602.
20. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M., RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci* **1998,** *38* (3), 511-22.
21. Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F.; Heeres, J.; Koymans, L. M.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A., SYNOPSIS: SYNthesize and OPtimize System in Silico. *J Med Chem* **2003,** *46* (13), 2765-73.
22. Meng, X. Y.; Zhang, H. X.; Mezei, M.; Cui, M., Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design* **2011,** *7* (2), 146-57.
23. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews. Drug discovery* **2004,** *3* (11), 935-49.
24. Aqvist, J.; Luzhkov, V. B.; Brandsdal, B. O., Ligand binding affinities from MD simulations. *Accounts of chemical research* **2002,** *35* (6), 358-65.

25. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of computer-aided molecular design* **2001,** *15* (5), 411-28.
26. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins* **2003,** *52* (4), 609-23.
27. Fuhrmann, J.; Rurainski, A.; Lenhof, H. P.; Neumann, D., A new Lamarckian genetic algorithm for flexible ligand-receptor docking. *Journal of computational chemistry* **2010,** *31* (9), 1911-8.
28. (a) Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *Journal of medicinal chemistry* **1999,** *42* (5), 791-804; (b) Ishchenko, A. V.; Shakhnovich, E. I., SMall Molecule Growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein-ligand interactions. *Journal of medicinal chemistry* **2002,** *45* (13), 2770-80.
29. (a) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P., Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of medicinal chemistry* **1999,** *42* (25), 5100-9; (b) Feher, M., Consensus scoring for protein-ligand interactions. *Drug discovery today* **2006,** *11* (9-10), 421-8.
30. Brady, G. P., Jr.; Stouten, P. F., Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des* **2000,** *14* (4), 383-401.
31. DeLano, W. L. *PyMol Molecular Graphic System*, version1.2r1; DeLano Scientific: South San Francisco,CA, 2009.
32. (a) Beccari, A. R.; Cavazzoni, C.; Beato, C.; Costantino, G., LiGen: a high performance workflow for chemistry driven de novo design. *Journal of chemical information and modeling* **2013,** *53* (6), 1518-27; (b) Beato, C.; Beccari, A. R.; Cavazzoni, C.; Lorenzi, S.; Costantino, G., Use of experimental design to optimize docking performance: the case of LiGenDock, the docking module of LiGen, a new de novo design program. *Journal of chemical information and modeling* **2013,** *53* (6), 1503-17; (c) Beato, C.; Beccari, A. R.; Cavazzoni, C.; Lorenzi, S.; Costantino, G., The use of experimental designs to optimize docking performances. The case of LiGenDock, the docking module of LiGen, a new de novo design program. . *Journal of Chemical Information and Modeling* **submitted**.
33. Korb, O.; Stutzle, T.; Exner, T. E., Empirical scoring functions for advanced protein-ligand docking with PLANTS. *Journal of chemical information and modeling* **2009,** *49* (1), 84-96.
34. Meng, E. C.; Shoichet, B. K.; Kuntz, I. D., Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992,** *13* (4), 505-524.
35. (a) Montes, M.; Miteva, M. A.; Villoutreix, B. O., Structure-based virtual ligand screening with LigandFit: pose prediction and enrichment of compound collections. *Proteins* **2007,** *68* (3), 712-25; (b) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M., LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of molecular graphics & modelling* **2003,** *21* (4), 289-307.
36. Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D., Molecular docking using shape descriptors. *J. Comput. Chem.* **1992,** *13* (3), 380-397.
37. Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L., 3rd, Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *Journal of computational chemistry* **2004,** *25* (2), 265-84.
38. Schneidman-Duhovny, D.; Nussinov, R.; Wolfson, H. J., Predicting molecular interactions in silico: II. Protein-protein and protein-drug docking. *Current medicinal chemistry* **2004,** *11* (1), 91-107.
39. (a) Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R., Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *Journal of chemical information and modeling* **2014,** *54* (6), 1700-16; (b) Li, Y.; Han, L.; Liu, Z.; Wang, R., Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. *Journal of chemical information and modeling* **2014,** *54* (6), 1717-36.
40. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *Journal of medicinal chemistry* **2006,** *49* (23), 6789-801.

41. Waszkowycz, B., Towards improving compound selection in structure-based virtual screening. *Drug discovery today* **2008,** *13* (5-6), 219-26.
42. Marcou, G.; Rognan, D., Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of chemical information and modeling* **2007,** *47* (1), 195-207.
43. (a) Bouvier, G.; Evrard-Todeschi, N.; Girault, J. P.; Bertho, G., Automatic clustering of docking poses in virtual screening process using self-organizing map. *Bioinformatics* **2010,** *26* (1), 53-60; (b) Mantsyzov, A. B.; Bouvier, G.; Evrard-Todeschi, N.; Bertho, G., Contact-based ligand-clustering approach for the identification of active compounds in virtual screening. *Advances and applications in bioinformatics and chemistry : AABC* **2012,** *5*, 61-79.
44. Rotstein, S. H.; Murcko, M. A., GroupBuild: a fragment-based method for de novo drug design. *J Med Chem* **1993,** *36* (12), 1700-10.
45. (a) Halgren, T. A., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996,** *17* (5-6), 490-519; (b) Halgren, T. A.; Nachbar, R. B., Merck molecular force field. IV. conformational energies and geometries for MMFF94. *Journal of Computational Chemistry* **1996,** *17* (5-6), 587-615; (c) Halgren, T. A., Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *Journal of Computational Chemistry* **1996,** *17* (5-6), 616-641; (d) Halgren, T. A., Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *Journal of Computational Chemistry* **1996,** *17* (5-6), 520-552; (e) Halgren, T. A., Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *Journal of Computational Chemistry* **1996,** *17* (5-6), 553-586.

Development and Validation Of In Silico Tools For Efficient Library Design and Data Analysis in High Throughput Screening Campaigns

Chapter 4: Targeting The Minor Pocket Of C5ar For The Rational Design Of An Oral Available Allosteric Inhibitor

4 TARGETING THE MINOR POCKET OF C5AR FOR THE RATIONAL DESIGN OF AN ORAL AVAILABLE ALLOSTERIC INHIBITOR

4.1 INTRODUCTION

Inflammatory and neuropathic pain are the most prevalent types of pathological pain and represent important health problems. Whereas inflammatory pain is one of the classical symptoms of the inflammatory process, neuropathic pain arises from any of multiple nerve lesions or diseases with symptoms including hyperalgesia or allodynia.[1,2] Some of the most powerful painkillers, including opioids and non-steroidal anti-inflammatory drugs, are only partially effective and prolonged exposure can cause unwanted effects.[3 4] As a result, there is continuous effort to identify novel therapeutics for pain control with alternative biological mechanisms and that elicit fewer side effects.

Inflammatory mediators, including cytokines/chemokines, play a critical role in the pathogenesis of inflammatory and neuropathic pain.[5,6] Emerging evidence suggests that C5a, the anaphylatoxin produced by complement activation, has potent nociceptive activity in several models of inflammatory and neuropathic pain by interacting with its selective receptor C5aR.[7 8] C5aR belongs to the class A subfamily of the seven-transmembrane G protein-coupled receptors (7TM GPCR)[9] and is widely expressed in immune cells, including neutrophils (PMN), monocytes, microglia, and in non-immune cells, including neurons in the central nervous system (CNS) and dorsal root ganglia (DRG).[10 11]

Evidence for a role of C5a in nociception sensitization has been obtained in several models of inflammatory pain. For instance, C5a was produced at the inflammatory sites and elicited mechanical hyperalgesia by activating the C5aR on infiltrated PMN.[7] Direct intraplantar injection of C5a in mice elicited both heat and mechanical hyperalgesia by sensitizing primary afferent C-nociceptors.[12 13] Local activation of C5aR has also been implicated in the pathogenesis of postsurgical pain, a model of postoperative pain.[13] Finally, local administration of PMX-53, a C5aR antagonist, attenuated mechanical hyperalgesia induced by carrageenan, zymosan, or lipopolysaccharide.[7] In addition to the peripheral role of C5a/C5aR in inflammatory pain, up-regulated levels of C5 and C5aR have been found in spinal cord microglia in animals subjected to spared nerve injury (SNI), a model of neuropathic pain.[8] Indeed, C5 null mice or the infusion of PMX-53 into the intrathecal space reduced neuropathic pain hypersensitivity in the SNI model.[8] Collectively, these data suggest that a neuroimmune interaction in the periphery and spinal cord through activation of the complement cascade and the production of C5a contributes to the genesis of both inflammatory and neuropathic pain.

As for other peptidergic GPCRs, the efforts to identify small molecular weight C5aR antagonists have led to a limited number of molecules, mostly lacking adequate potency and selectivity.[14] The most promising candidate so far described, PMX53, is a cyclic peptidomimetic antagonist, designed to mimic the C-terminal portion of C5a.[15] Despite the encouraging results obtained in preclinical studies, as with many peptide drugs, the development of PMX53 has been limited by its short half-life and unfavorable bioavailability.[16] In the present study, we report the successful design of a novel non-peptidic C5a allosteric small molecular weight (SMW) inhibitor driven by the structural information on a minor pocket spanning between TM1, 2, 3, 6 and7 that is highly conserved across the GPCR family and that has been recently proposed as a key motif for the intracellular activation process Reparixin was previously reported as a neutral allosteric inhibitor of CXCR1 and CXCR2 that binds the TM in a region that overlaps the minor pocket.[17 18] Combining the information from independent sources on structural and functional features of allosteric sites in

homologous chemokine receptors, this paper intends to provide the first example of *de novo* design of a new class of allosteric SMW inhibitors of a GPCR not belonging to the chemokine receptors family, C5aR. The preclinical candidate, DF2593A, is a potent and orally active C5a noncompetitive allosteric inhibitor with significant antinociceptive effects in a wide range of inflammatory and neuropathic pain models.

4.2 RESULTS

4.2.1 Binding mode characterization of DF2593A to C5aR

The human C5aR (hC5aR) homology model was originally built using the human CXCR1-reparixin complex [19] as a template and subsequently refined and compared with the C5aR model built starting from the human CCR5 (hCCR5) crystal structure (PDB code: 4MBS), in which CCR5 is bound with the marketed HIV allosteric drug maraviroc.[20] [21] Sequence identity between hCCR5 and hC5aR is 21.3% while sequence similarity is 52.4%. Despite a low sequence identity, the key structural features defining the minor pocket, the proline kink in TM2 and the water-mediated hydrogen bond network between the intracellular segments of TM1, 2, 3, 6 and 7, is highly conserved between chemokine receptors and C5aR. With the aim of developing a specific site-binding model in C5aR, the pattern of polar interactions involved in the anchorage of reparixin at CXCR1 was analyzed in detail.[19] Two out of the three key residues of TM 3, 6 and 7 (Asn120$^{3.35}$, Tyr258$^{6.51}$ and Glu291$^{7.39}$) are fully conserved (Asn119$^{3.35}$ and Tyr258$^{6.51}$) in C5aR,[19] [22] and the glutamic residue 7.39 (Val284$^{7.39}$ in C5aR) is replaced by Asp282$^{7.35}$ one helix turn above (Figure 28A and Table 2).

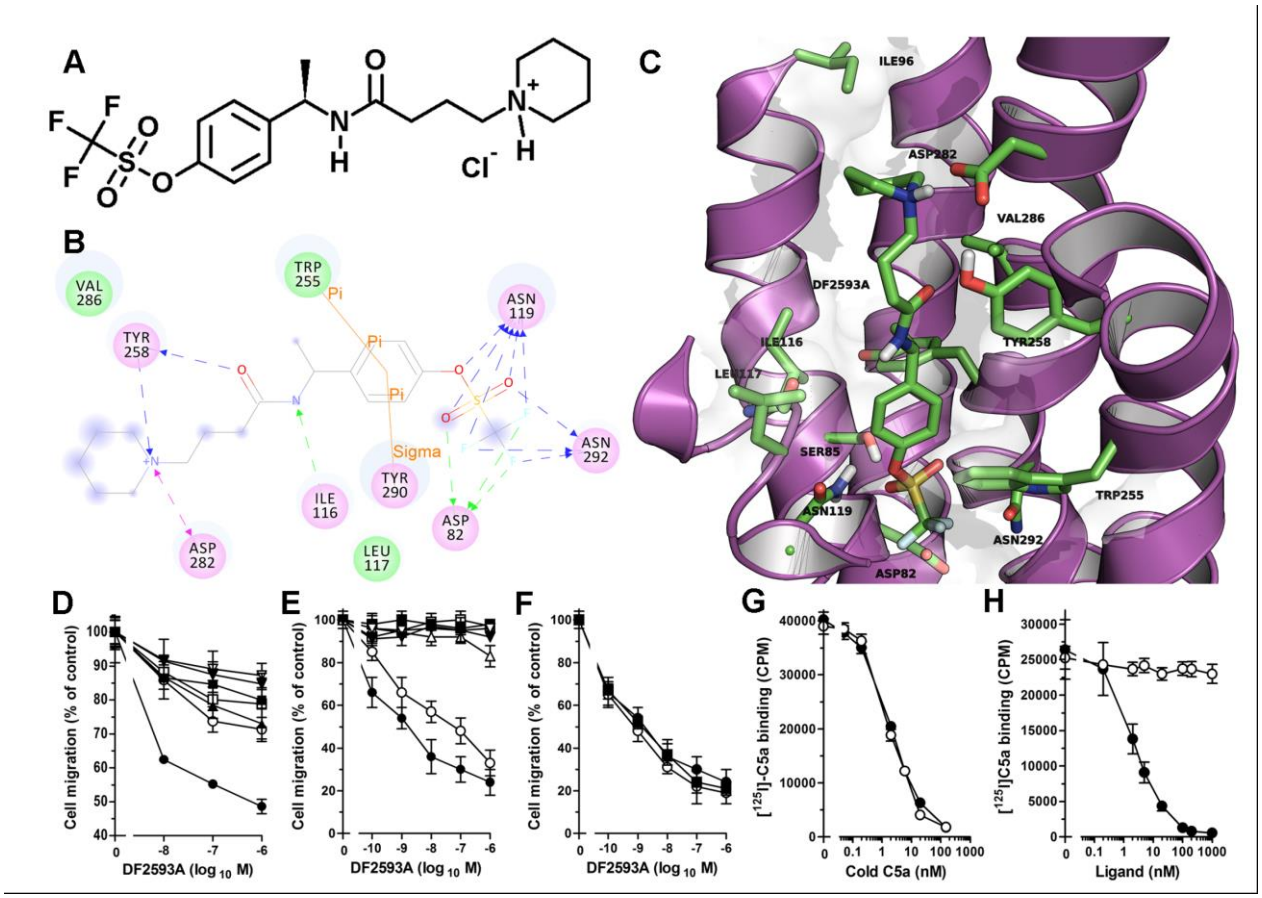


Figure 28. ***DF2593A is a selective noncompetitive allosteric inhibitor of C5aR.*** *(A) 2D structure of DF2593A. (B) 2D representation of DF2593A binding mode inside C5aR. Pink, green/blue and orange lines represent respectively: charge-charge interaction, H-bond/Halogen-bond interaction, Pi-Pi and sigma interactions. (C) 3D representation of DF2593A in the TM region of C5aR (violet). Interacting residues are shown using the following color scheme: C(green), N(blue), O(red), S(yellow), F(cyan) and H(white); labeling*

*is based on Ballesteros-Weinstein numbering scheme. (D) Effect of DF2593A on the C5a-induced migration of wt (●), L41A (○), F93A (□), I96A (▲), L278A (▼), N119A (▽), and Y258A (■) C5aR/L1.2 transfectants preincubated with indicated DF2593A concentrations and then stimulated with 10 nM of C5a for 4 h. All mutants show a significant (p>0.05) resistance to DF2593A at all doses tested as compared to wild-type C5aR transfectants (E) Effect of DF2593A on the C5a-induced migration of human PMN preincubated for 15 minutes with indicated DF2593A concentrations and then stimulated with 10 nM C5a (●), 1 μM C5a-desarg (○), 10 nM CXCL1 (□), 1 nM CXCL8 (▲), 10 nM CXCL12 (△), 30 nM CCL3 (▼), and 10 nM fMLP (▼) for 1 h. (F) Effect of DF2593A on the C5a-induced migration of human (■), mouse (○) and rat (●) PMN preincubated with indicated DF2593A concentrations and then stimulated with 10 nM C5a of corresponding origin. (G) Human PMN incubation with vehicle (●) or 1 μM of DF2593A (○) and then aliquots of 0.2 nM of [^{125}I]-C5a and serial dilution of cold C5a were added to 10^6 cells in 100 mL of binding medium. (H) Human PMN were directly exposed to vehicle or different concentrations of DF2593A (○) and then aliquots of 0.2 nM of [^{125}I]-C5a and vehicle (DF2593A-treated PMN) or serial dilution of unlabeled C5a (vehicle-treated PMN) (●) were added to 10^6 cells in 100 mL of binding medium. In panels D to F data are expressed as % of migrated cells observed in the absence of DF2593A (mean ± S.D. of three independent experiments). **P < 0.01 versus cell migration in the absence of DF2593A by Mann–Whitney U-test. In panels G and H data are reported as mean ± S.D. of three independent experiments.*

Table 2

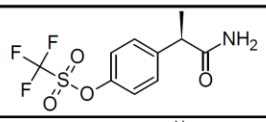
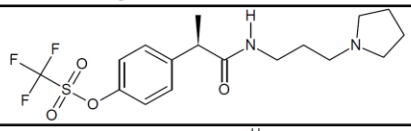
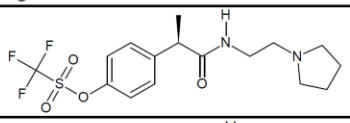
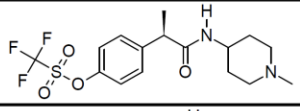
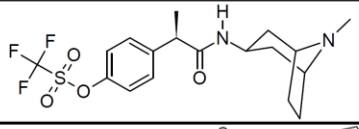
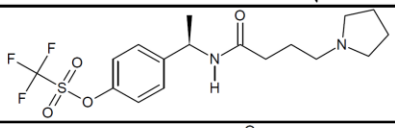
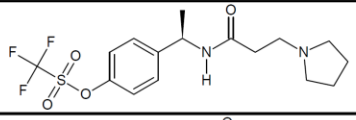
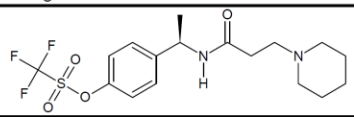
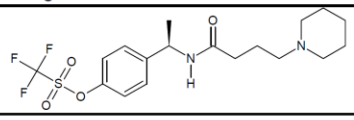
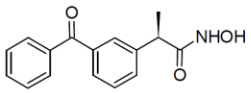
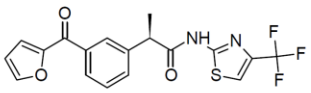
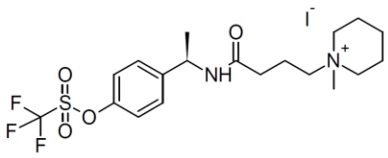
Receptor (Uniprot ID)	Pocket polar residues			
	3.35	**6.51**	**7.35**	**7.39**
CXCR1 (P25024)	**N120**	**Y258**	**L287**	**E291**
CXCR2 (P25025)	**N129**	**Y267**	**L296**	**E300**
C5AR (P21730)	**N119**	**Y258**	**D282**	**V286**

Sequence alignment of key polar residues among CXCR1/2 and C5aR.Two out of the three key polar residues (Asn1203.35, Tyr2586.51 and Glu2917.39) of CXCR1 are conservatively replaced (Asn1193.35 and Tyr2586.51) in C5aR, whereas the lack of the glutamate in position 7.39 (Val2847.39 in C5aR) is compensated by Asp2827.35. Uniprot ID codes are provided in brackets. Amino acids are color coded by type: red, acidic residues; green, hydrophobic aliphatic residues; blue, aromatic residues; violet, non-charged polar residues.

Furthermore, as for CXCR1, the minor pocket in C5aR is surrounded by a large cluster of hydrophobic residues (Leu41$^{1.36}$, Phe93$^{2.61}$, Ile96$^{2.64}$ and Leu278$^{7.31}$) in the upper region of the TM layer (Fig. 1C). Interestingly, Asp82$^{2.50}$, the crucial amino acid in the conserved motif of TM2: $L_{92\%}XXXD_{91\%}$ is also present in the binding cavity of C5aR.[23] The formation of a hydrogen bond between Asp82$^{2.50}$ and Asn296$^{7.49}$ in TM7 is deemed crucial to activate C5aR.[24] Accordingly, the freezing of Asp82$^{2.50}$ in the inactive position with DF2593A may prevent the crucial movement to switch C5aR to its active state and it may represent a mechanistic explanation for the inhibitory effect of DF2593A.

Based on these structural characteristics, a site-targeting library was designed to identify a new class of allosteric C5aR modulators. The phenylpropionic moiety was chosen as a privileged scaffold to accommodate the shallow crevice between TM2, 3, 6 and 7 and the trifluoromethyl group in the para position was selected as a promising substituent to establish a bidentate interaction with Asn119$^{3.35}$ and Asp82$^{2.50}$ bridging TM3 and TM2 associated with the feature of the trifluoromethyl portion to engage additional hydrophobic interactions [25] [26] Table 3, **1**).

Table 3.

Cmpd	Structure	CXCL8[a]	CXCL1[b]	C5a[c]
1		8 ± 0.6	26 ± 2.3	>10.000
2		5 ± 0.2	20 ± 1.1	50 ± 3.4
3		10 ± 0.3	30 ± 2.2	60 ± 4.2
4		>10.000	>10.000	50 ± 3.2
5		>10.000	>10.000	60 ± 1.2
6		>10.000	>10.000	40 ± 2.3
7		>10.000	>10.000	300 ± 19
8		>10.000	>10.000	150 ± 10
9		>10.000	>10.000	10 ± 1.3
10		>10.000	>10.000	10 ± 1.1
11		>10.000	>10.000	10 ± 0.8
12		>10.000	>10.000	10 ± 0.6

Seeking the key polar interactions with Asp282$^{7.35}$ and Tyr258$^{6.51}$, a series of carboxamide and aminocarbonyl residues bearing basic groups were synthesized. Flexible tertiary ω-aminoalkylamides (Table 3, **2-3**) were Found to be dual significant inhibitors of C5a and CXCL8 induced chemotaxis whereas the insertion of conformationally rigid cyclic structures led to a full selectivity versus the C5aR (Table 3, **4-5**).

The geometry and the electronic status of the amido group were previously demonstrated [25] to be key determinants in establishing the polar network in CXCR1/2 binding sites. Thus the simple inversion of the

amide group in compound **1** was sufficient to lose affinity at CXCR1/2 leading to the fully C5aR selective isomer **6**. The optimization of the alkyl chain length and hydrophobicity (Table 3, **6-9)** led to the selection of DF2593A (Table 3, **9**) as the most promising lead for further *in vitro* and *in vivo* characterization (Figure 28B and Figure 28C). Extensive molecular dynamics and automatic docking runs on the C5aR/DF2593A complex were performed to generate a reliable binding mode hypothesis. Based on the refined model, DF2593A is able to bind C5aR by two principal polar interaction patterns: (1) charge-charge interaction between Asp282$^{7.35}$and the charged nitrogen of the piperidine ring of DF2593A and (2) a network of polar interactions between the triflate and the couple Asn119$^{3.35}$/Asp82$^{2.50}$. Intriguingly, halogen bond interactions are engaged between fluorine atoms and carbonyl oxygen of both Asn119$^{3.35}$ and Asp82$^{2.50}$.

The polar network is reinforced by the hydrogen bond between the carbonyl group of Ile116$^{3.32}$ and the amide moiety of DF2593A. In addition, the phenyl ring of DF2593A is involved in a pi-pi interaction with Tyr290$^{7.43}$. As mentioned, the binding cavity is also surrounded by hydrophobic residues, like Leu41$^{1.36}$, Phe93$^{2.61}$, Ile96$^{2.64}$, Leu278$^{7.31}$ and Val286$^{7.39}$ (Figure 28). It has been observed that a Val286Ala$^{7.39}$ mutant led to a moderately inactive C5aR supporting the hypothesis of a prominent functional role of TM7.[27]

The allosteric nature of the C5aR minor pocket [17] and the inhibitory activity of DF2593A were then investigated. Wild-type receptor and seven Ala-scanning mutants of C5aR were transiently transfected in L1.2 cells and tested for functionality. All mutants but Asp282$^{7.35}$ were non significantly impaired in their cell surface expression (see Fig. S2) and sustained effective cell migration in response to C5a (see Fig. S3).

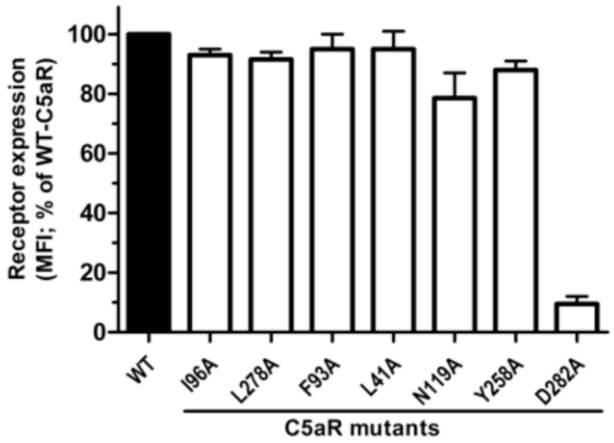


Figure 29. Expression levels of C5aR mutants. Cell surface expression of wild-type (black column) and indicated point mutants (while columns) of C5aR transiently transfected in L1.2 cells. Results are mean ± SEM of three independent experiments. D282A-C5aR is the only mutant significantly impaired as compared to wild-type C5aR (p<0.01).

Transfectants were then tested in the C5a-induced chemotaxis assay in the presence of increasing concentrations of DF2593A. As shown in Figure 28D, the Asn119$^{3.35}$Ala mutant is fully resistant to inhibition by DF2593A, supporting the hypothesis that Asn119$^{3.35}$ is a key residue for ligand recognition. The dramatic effect of Tyr258$^{6.51}$Ala mutant was coherent with the ligand/receptor complex in which the interaction established by the phenol group is essential to orientate Asp282$^{7.35}$ towards the piperidine nitrogen of DF2593A. Tyr258$^{6.51}$ hydroxyl group could also act as a hydrogen bond donor versus the carbonyl group in the central amide moiety of DF2593A. According to the proposed model, also the bulky hydrophobic Leu41$^{1.36}$, Phe93$^{2.61}$, Ile96$^{2.64}$Ala and Leu278$^{7.31}$, even though not directly involved in ligand binding, are structural and energetic determinants of the hydrophobic domain governing the pattern of interactions within the hydrophobic cluster. The replacement of these residues with Ala profoundly alters the shape and size of the pocket and coherently significantly compromises DF2593A potency. Notably, sequence alignment confirmed that key residues involved in the putative DF2593A binding mode are well conserved

in both rat and mouse orthologs supporting the potential of the candidate for further preclinical pharmacology studies (Figure 30).

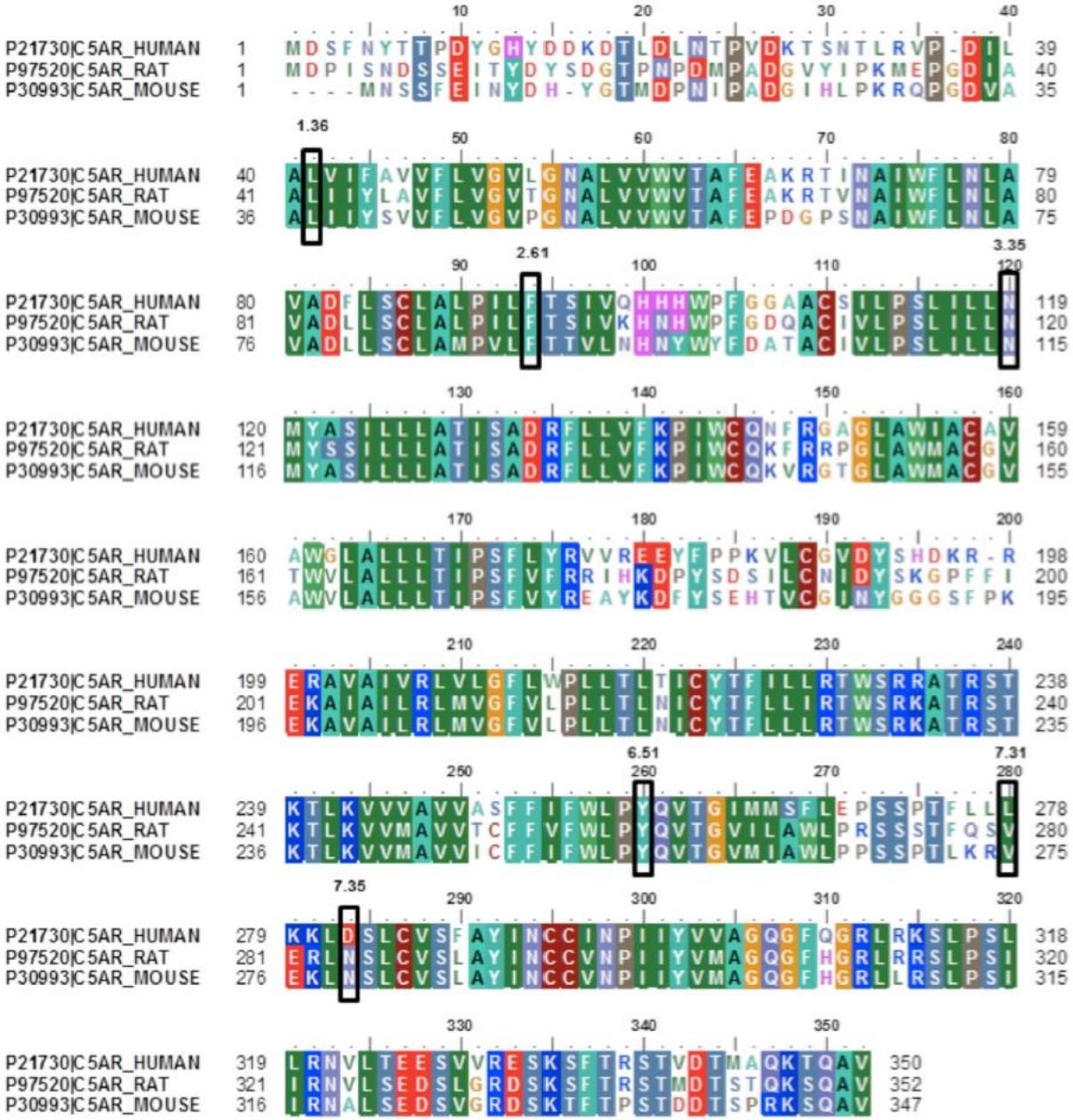


Figure 30. *Sequence alignment of human, mouse and rat C5aR. Asn1203.35 and Tyr2586.51 are fully conservatively in the C5aR orthologs, whereas Asp2827.35 is replaced by Asn7.35 in both the orthologs. Hydrophobic residues, Leu411.36 and Phe932.61 are also fully conserved in mouse and rat C5aR. Ile962.64 is conserved in rat (Ile972.64) and is replaced by Val932.64 in mouse C5aR. Leu278 7.31 is replaced by a similar hydrophobic residue, Val2767.31 and Val2807.31, in mouse and rat respectively. Key residues of the allosteric pocket are shown in black boxes with Weinstein Ballesteros annotation. Uniprot ID codes are provided.*

4.2.2 In vitro characterization of DF2593A

Pharmacological characterization showed that DF2593A did not inhibit spontaneous cell migration *per se* and was > 1000-fold selective versus other chemoattractants, including CXCL8 and CXCL1 (IC_{50} > 10 μM) (Fig. 1E). DF2593A effectively inhibited C5a-induced human PMN migration (IC_{50} = 5.0 nM) and cross-reacted with rat and mouse orthologs (IC_{50} = 6.0 nM and IC_{50}= 1.0 nM, respectively) (Figure 28F). DF2593A (tested

at 10 μM) was selective on a panel of different GPCRs and ion channels (Table 4.) and was also completely inactive (% inhibition = 10 ± 3; mean ± SD; three experiments performed in duplicate) in a PGE_2 production assay.[28]

Table 4.

GPCR	Ligand	Mean±SD
α_{1A}-adrenoceptor	[^{125}I]-HEAT	105.0±5.7
α_{2A}-adrenoceptor	[^{3}H]-Rawolscine	88.5±2.8
β1-adrenoceptor	[^{3}H]-CGP12177	95.3±14.2
β_2-adrenoceptor	[^{3}H]-CGP12177	97.3±5.9
B_1	[^{3}H]-(des-Arg^{10},Leu^{9})-Kallidin	95.5±6.9
B_2	Bradykinin	98.5±6.0
CB1	[^{3}H]-CP-55,940	92.5±5.8
CB2	[^{3}H]-CP-55,940	88.5±13.4
D2	[^{3}H]- Methylspiperone	91.7±1.0
D3	[^{125}I]-7-OH-PIPAT	85.4±2.3
H1	[^{3}H]-Pyrilamine	62.7±0.7
H2	[^{125}I]-Aminopotentidine	98.0±9.31
M2[a]	[^{3}H]-N-methyl scopolamine	10.1±0.8 (10 μM)
		60.1±1.1 (1 μM)
		100.1±5.1 (0.1 μM)
M3[a]	[^{3}H]-N-methyl scopolamine	15.3±0.7 (10 μM)
		55.1±3.4 (1 μM)
		95.4±3.0 (0.1 μM)
NK_1	[^{125}I]-Substance P	92.5±0.7
μ	[^{3}H]-DAMGO	90.4±3.1
κ	[^{3}H]-U69,593	88.4±3.0
δ	[^{3}H]-Naltrindole	94.1±14.9
NOP	[^{3}H]-Nociceptin	110.6±2.9
5-HT1A	[^{3}H]-8-Hydroxy-DPAT	97.3±17.8

Ion channel	Ligand	IC_{50} (μM)
TRPM8	Icilin	> 10
TRPV1	Capsaicin	> 10
TRPV4	GSK1016790A	> 10

Data are presented as mean (± SD, n = 3) of the % of binding of radioligand after the incubation with DF2593A at the single concentration of 10 μM. For TRPM8, TRPV1 and TRPV4 experiments, data are presented as IC_{50} values of three different experiments. [a] The effect of DF2593A was evaluated at 10, 1 and 0.1 uM. DF2593A retains about 1000 fold of selectivity versus M2 and M3.

Binding experiments carried out on PMN membranes with radiolabelled C5a showed that DF2593A did not compete with binding of C5a. Pretreatment of PMN with DF2593A (1 μM) did not change C5a affinity for C5aR (K_d = 1.8 ± 1 nM and 1.0 ± 0.4 nM in vehicle and DF2593A-treated groups, respectively; n=3/group). In

addition, DF2593A did not affect the number of C5aR molecules expressed on the cell membrane (127000 ± 13000 and 133000 ± 21000 binding sites/cell in vehicle and DF2593A-treated groups, respectively) and did not alter C5a binding to C5aR in displacement experiments (Figure 28G and Figure 28H).

4.2.3 Pharmacokinetic profile of DF2593A

The pharmacokinetics of DF2593A was investigated in mice after intravenous and oral administration. Following oral administration (1 mg/kg) in mice, DF2593A was well absorbed (F = 83%) with a C_{max} of 0.1 μM, a t_{max} of 1.2 h and plasma samples harvested over a period of 12 h, generated free plasma drug concentrations in the range 0.016 - 0.004 μM, about 4 to 16-fold greater than its *in vitro* IC_{50}. For brain permeation studies, DF2593A was dosed i.v. at 10 mg/kg and the levels of DF2593A were 913 ng/g tissue and 673 ng/ml in brain and plasma respectively, at 5 min post dose. At 8 h post dose, significant levels of DF2593A were still present in the brain (114 ng/g) and also in the plasma (56 ng/ml). The brain-to-plasma ratios of DF 2593A ranged from 1.36 (at 5 min post dose) to 2.04 at 8 h post dose. In parallel, the distribution coefficient, logBB, increased from 0.13 (at 5 min) to 0.3 (at 8 h), suggesting good permeation of the molecule into the brain.

4.2.4 Antinociceptive effects of DF2593A in several models of inflammatory pain

C5a/C5aR interactions mediate carrageenan-induced mechanical hyperalgesia in rodents, suggesting it was a reasonable model to evaluate the potential analgesic effect of DF2593A.[7] In corroboration, carrageenan-induced inflammatory hyperalgesia is reduced in $C5aR^{-/-}$ mice compared with WT mice (Figure 31 (S4)).

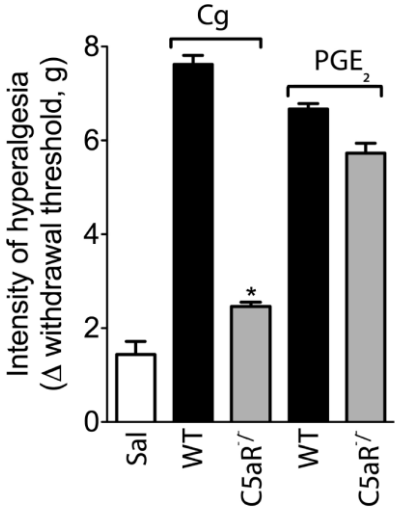


*Figure 31. Inflammatory hyperalgesia induced by carrageenin (Cg), but not by PGE2, is reduced in C5a null mice. WT or C5aR-/- mice received intraplantar injection of Cg (100 μg/paw), PGE2 (100 ng/paw), or saline. At 3 h mechanical hyperalgesia was determined. Data are shown as mean ± SEM. * P<0.05when compared to wild type group; n=6.*

Consistently with its *in vitro* potency, oral administration of DF2593A inhibited carrageenan-induced mechanical hyperalgesia in a dose-dependent manner, but did not alter the mechanical threshold in naïve mice (Figure 32A Fig. 2A). C5a-induced hyperalgesia was also blocked by DF2593A, supporting the concept that DF2593A blocks hyperalgesia by inhibiting the C5aR (Figure 32 Fig. 2B). Neither DF2593A nor $C5aR^{-/-}$ deletion affected mechanical hyperalgesia induced by PGE_2 injected in the mice paws (Figure 32C and Figure 32D). Furthermore, DF2593A did not alter epinephrine-induced hyperalgesia in the mice paws (Figure 32Fig. 2C). In an attempt to test the effect of DF2593A against chemical nociception, a writhing test induced by zymosan was studied.[29] Oral treatment of mice with DF2593A significantly reduced zymosan-induced writhing response (Figure 32D). DF2593A had no opioid-like affect in the hot-plate test and did not

cause any motor impairment that could have accounted for the analgesic effects observed (Figure 32E and F).

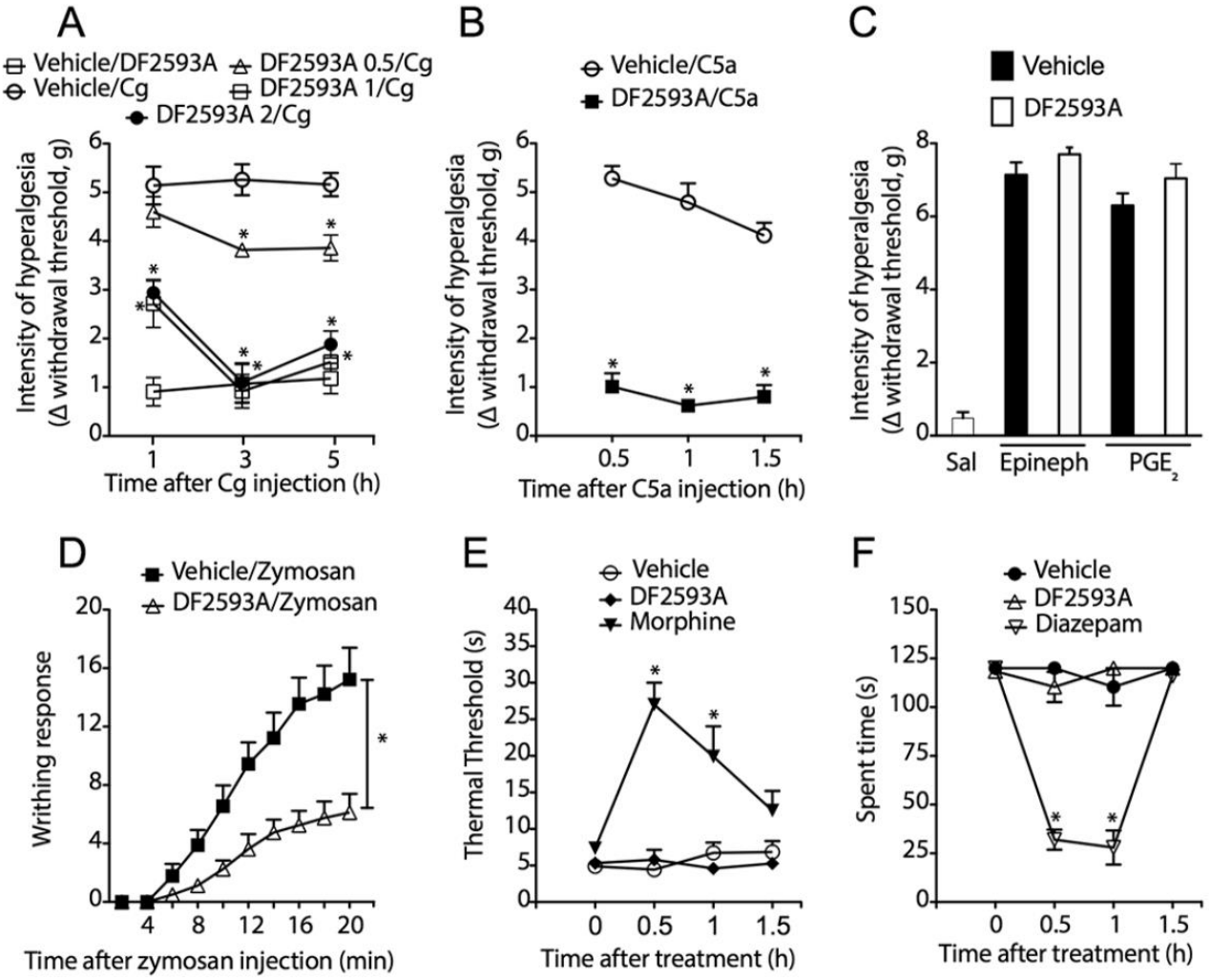


*Figure 32. Antinociceptive effects of DF2593A on acute inflammatory pain models. (A) Mice were treated orally with vehicle or DF2593A (0.25 - 2 mg/Kg) 50 minutes before intraplantar (i.pl) injection of Carrageenan (Cg, 100 µg/paw) or vehicle (saline). (B) Mice were treated orally with vehicle (saline) or DF2593A (1 mg/Kg), 50 minutes before i.pl injection of C5a (300 ng/paw). (C) Mice were treated orally with vehicle (saline) or DF2593A (1 mg/Kg), 50 minutes before i.pl injection of PGE2 (100 ng/paw) or epinephrine (300 ng/paw). Mechanical hyperalgesia was evaluated 60 min after stimuli injection. (D) Mice were treated orally with vehicle (saline) or DF2593A (1 mg/Kg) 50 minutes before intraperitoneal injection of zymosan (1 mg/cavity). The number of writhing response induced by zymosan was evaluated during 20 min n=8. (E) The thermal nociceptive threshold in naïve mice was evaluated before and at indicated time after DF2593A treatment (1 mg/kg) using hot-plate test. Morphine (8 mg/kg s.c.) was used as a positive control. (F) Naïve mice were treated orally with vehicle or DF2593A (1 mg/Kg) followed at indicated times by evaluation in rota-rod test. Diazepam (7.5 mg/kg, s.c) was used as a positive control. Data are shown as mean ± SEM. (n = 6-8 per group). *P<0.05 as compared with vehicle-treated animals.*

Treatment with DF2593A or genetic deletion of C5aR (C5aR$^{-/-}$ mice) did not alter carrageenan-induced PMN accumulation (Figure 33), in agreement with our previous findings.[7]

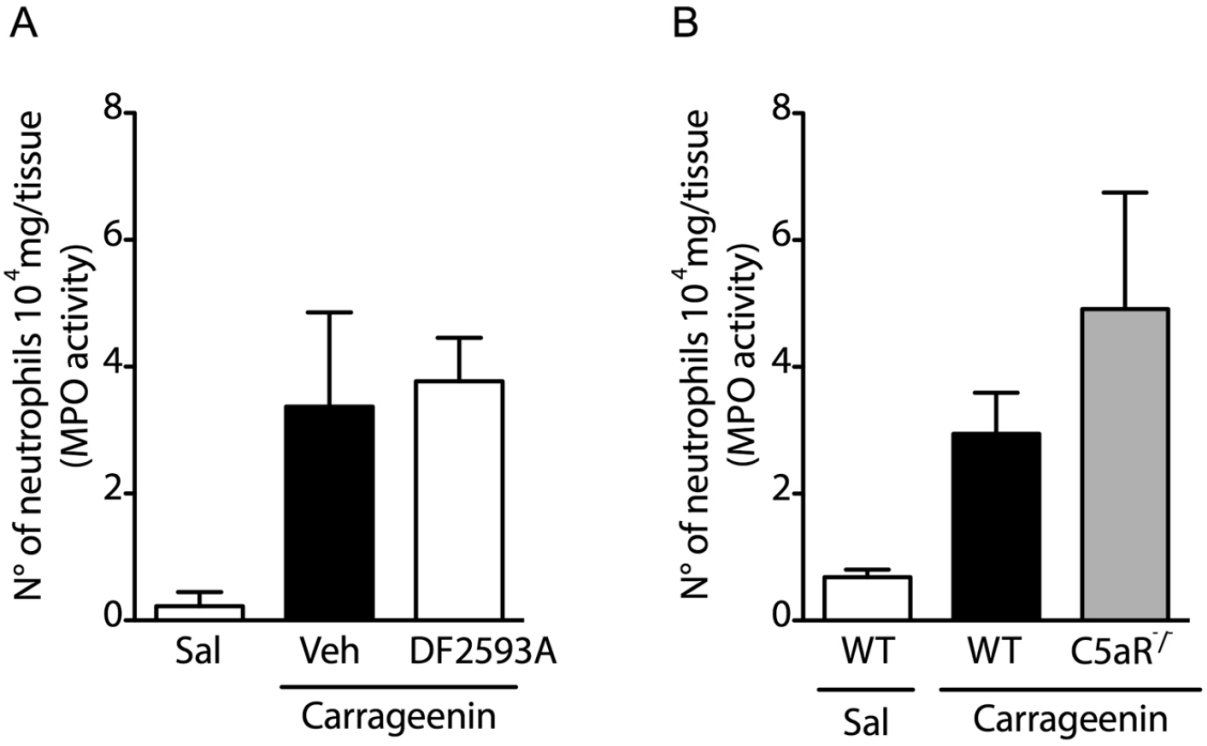


Figure 33. Role of C5a in neutrophil migration during carrageenin (Cg)-induced inflammation in mice paw. (A) Effect of DF2593A on the Cg-induced increase in paw MPO activity. Mice were treated orally with vehicle (saline) or DF2593A (1 mg/Kg) 50 min before intraplantar injection of Cg (100 µg/paw). MPO activity (number of neutrophils 104/mg tissue) was evaluated in the plantar tissue 5 h after Cg injection (B) WT and C5aR-/- mice receive an intraplantar injection of Cg (100 µg/paw). MPO activity (number of neutrophils 104/mg tissue) was evaluated in the plantar tissue 3 h after Cg injection. Data are shown as mean ± SEM. n=6.

The analgesic effect of DF2593A was then tested in the complete Freund's adjuvant (CFA)-induced chronic inflammatory pain model. Oral pretreatment with DF2593A inhibited CFA-induced mechanical hyperalgesia (Figure 34A). Coherently with its plasma levels, the effect of DF2593A was detected at least 6 h after CFA injection, but not at 24 h (Figure 34A). We then studied whether the post treatment with DF2593A could modify mechanical CFA-induced hyperalgesia and whether the analgesic effect was maintained. As reported in Fig. 3B and C, delayed treatment with DF2593A (24 h after CFA injection) reduced CFA-induced mechanical and thermal inflammatory hyperalgesia. In a therapeutic setting, treatment with DF2593A starting at 24 h after CFA injection and then twice a day for one week reduced mechanical hyperalgesia (Figure 34D). Interestingly, the effect of DF2593A was maintained and when DF2593A treatment was stopped, mechanical hyperalgesia resumed to basal levels (Figure 34D). There is evidence to suggest that C5a contributes to the pathogenesis of rheumatoid arthritis, including arthritic pain.[7 30] Therefore, in the next step we evaluated the effect of DF2593A in the genesis of articular hyperalgesia in two models of arthritis in mice, antigen- and zymosan-induced arthritis. Oral pretreatment with DF2593A effectively inhibited articular hyperalgesia in both models (Figure 34E and F).

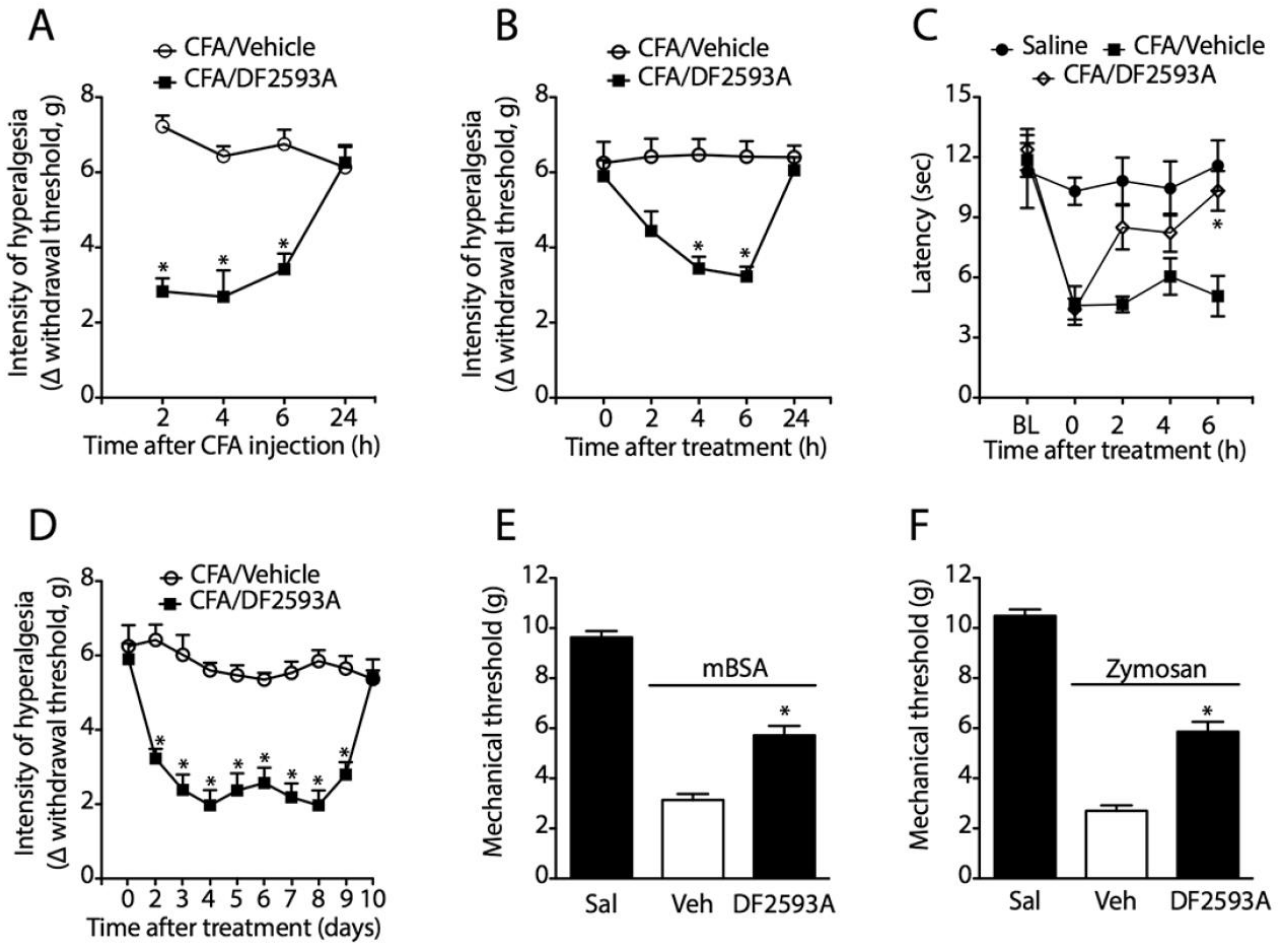


*Figure 34. Effects of DF2593A on CFA-induced mechanical and thermal hyperalgesia and arthritic nociception. (A) Mice were treated orally with vehicle or DF2593A (1 mg/Kg) 50 minutes before i.pl injection of CFA (10 µL/paw). (B) Mice were treated orally with vehicle or DF2593A (1 mg/Kg) 24 h after i.pl injection of CFA (10 µL/paw). (B) Mechanical or (C) thermal nociceptive threshold was evaluated 2-24 h after DF2593A treatment. (D) DF2593A (1 mg/kg) or vehicle was given at 24 h after CFA injection and then twice a day until 7 days after CFA. Mechanical hyperalgesia was evaluated at 6 h after the first daily dose of DF2593A. (E, F) Mice were pretreated orally with vehicle (saline) or DF2593A (1 mg/Kg) 50 minutes before intraarticular injection of mBSA (30 µg/joint) in immunized mice or before intraarticular injection of zymosan (150 µg/joint) in naive mice. Mechanical nociceptive threshold was evaluated 6 h after stimuli injection. Data are shown as mean ± SEM. (n = 6 per group).*P<0.05 as compared with vehicle-treated animals.*

4.2.5 Antinociceptive effects of DF2593A in the SNI model of neuropathic pain

As activation of the complement system at the site of nerve injury and in the spinal cord contributes to induction and establishment of neuropathic pain,[8 31] the effect of DF2593A in the SNI-induced neuropathic pain model in mice was evaluated. Oral treatment with DF2593A seven days after surgery clearly reduced mechanical hypersensitivity induced by SNI (Figure 35A). The effects of DF2593A persisted at least 6 h after treatment and, coherently with the pharmacokinetic profile, a progressive loss of antinociceptive effect was observed after 24 h (Figure 35A). Confirming the involvement of C5aR in the genesis of neuropathic pain, mechanical hypersensitivity 7 days after SNI was also reduced in C5aR-/- mice when compared with WT mice (Fig. 4B). Furthermore, treatment of $C5aR^{-/-}$ mice with DF2593A did not produce any further antinociceptive effect when compared with $C5aR^{-/-}$ mice treated with vehicle (Figure 35B). The mechanical nociceptive threshold of $C5aR^{-/-}$ naïve mice did not differ from WT mice (Mechanical nociceptive threshold of naïve WT mice: 7.2 ± 0.6 g and naïve $C5aR^{-/-}$ mice: 6.8 ± 0.1 g).

In a therapeutic setting, DF2593A was given twice a day from day 7 to day 14 after surgery and mechanical hypersensitivity measured 6 h after the first daily dose. As seen in Figure 35C, the effects of DF2593A were immediate and maintained over the observation period. Mechanical hypersensitivity returned to the same level of SNI control group two days after the suspension of the DF2593A treatment (Figure 35C).

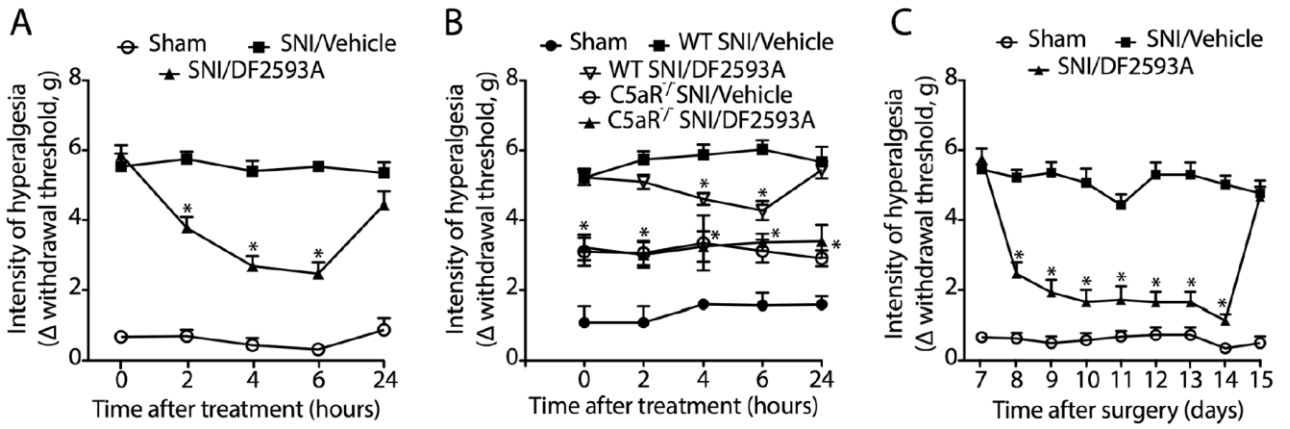


*Figure 35. Antinociceptive effects of DF2593A in the spared nerve injury model of neuropathic pain. (A) DF2593A (1 mg/kg) or vehicle (saline) were given orally 7 days after SNI surgery in mice. Mechanical hypersensitivity was evaluated from 2 to 24 h after oral administration of DF2593A. (B) Mechanical hypersensitivity was evaluated 7 days after SNI surgery in WT and C5aR-/- mice (time 0) followed by the treatment with vehicle or DF2593A (1 mg/kg). Mechanical hypersensitivity was then evaluated 2-24 h after DF2593A or vehicle treatment. (C) DF2593A (1 mg/kg per dose) or vehicle (saline) were given at day 7 and then twice a day until 14 days after SNI. Mechanical hypersensitivity was evaluated at 6 h after the first daily dose of DF2593A during one week. Data are shown as mean ± SEM. (n = 6 per group). *P <0.05 as compared with vehicle-treated animals.*

4.3 DISCUSSION

This paper reports the molecular conception, synthesis and characterization of the preclinical candidate DF2593A that shows potent and selective inhibitory effect on the C5a-induced PMN migration and optimal pharmacokinetic and pharmacological profile in a panel of relevant inflammatory and neuropathic pain experimental models.

In recent years, allosteric modulation of GPCRs has been proposed as a promising new paradigm for the design of potent and selective drugs with improved drug-like properties, finely modulating the receptor function. There is emerging evidence to suggest that, despite marked differences between the natural ligands, GPCRs share common activation mechanisms involving specific microswitches that regulate interhelical movements which offer unprecedented opportunities for the rational drug design of novel allosteric modulators. In this context, modeling and crystallographic studies of the GPCRs have identified in the TM region of GPCRs a major pocket and a minor pocket which has been proposed as a "triggering domain" not involved in natural ligand binding but crucial for the fine tuning of the global receptor activation process.[17] Previous studies elucidated that reparixin, an allosteric inhibitor of CXCL8 receptors, binds CXCR1 within the minor pocket locking the receptor in the inactive state by interhelical polar interactions with residues of TM1, TM3, TM6 and TM7.[19]

Guided by the hypothesis that this minor pocket may represent a functionally conserved site across the GPCR family, homology modeling studies and molecular dynamics simulations of C5aR were carried out. Despite the low overall identity between C5aR and chemokine receptors, sequence analysis revealed that the three key residues identified for reparixin and CXCR1 are conserved in C5aR.[22] Taking advantage of the structural diversity of other residues surrounding the minor pocket in C5aR and CXCR1, rational design was

addressed by targeting a specific pattern of interactions to retain full selectivity to C5aR. Extensive site-directed mutagenesis studies in C5aR confirmed that this region is not directly involved in ligand receptor binding or critical for receptor activation and function. This finding nicely fits with the most accredited model of C5a/C5aR interaction [31] according to which the flexible C terminal of C5a is implicated in the recognition of a domain distinct from the allosteric pocket and delimited mainly by Arg206$^{5.42}$, TM4, and the second extracellular loop. Nevertheless our studies clearly demonstrated that specific ligands at this minor but conserved site are potent inhibitors of the C5a-induced chemotaxis acting as neutral non-competitive allosteric inhibitors. Interestingly, as the key residues involved are well conserved in both rat and mouse orthologs, the above results pave the way for the rational design of allosteric C5aR inhibitors with most favorable cross-species reactivity characteristics.

Previous studies have characterized the role of the C5a/C5aR signaling in the genesis of inflammatory and neuropathic pain.[7] [8] Neuropathic pain is an important and relatively common clinical condition and available therapies are only partially effective. Among the events implicated in the genesis of neuropathic pain, neuroimmune interactions, including spinal activation of glial cells and the production of pro-inflammatory mediators, seem to be important for pain amplification. It was shown that expression of common genes, including the complement system genes, occurred in three different models of peripheral neuropathy.[8] These findings are in agreement with the inhibition of complement activation by the intrathecal administration of a soluble human complement receptor type 1 (sCR1), which prevented the activation of C3 and C5 convertases and reduced mechanical allodynia in various neuropathic pain models.[31] Additionally, C6 deficient rats still presented nerve injury-induced mechanical allodynia, suggesting that membrane attack complex assembling is not necessary for the induction of neuropathic pain states.[8] Accordingly, the nociceptive activity of the complement system appears to be strongly dependent on the C5a/C5aR interaction. Indeed, after nerve injury, expression of C5aR and C5 in the microglial cells of the spinal cord increased and C5 deficient mice or intrathecal treatment with the classical antagonist PMX-53, ameliorated nerve injury-induced allodynia.[8] Altogether, this preclinical evidence strongly supports the hypothesis that C5aR is an interesting target for neuropathic pain control.

The precise cellular site of action of DF2593A in the nociceptive system is not fully understood. Considering the aforementioned evidence of a role of C5a in the CNS, the efficacy of DF2593A in the SNI model may arise in part from its permeability to the blood-brain barrier. However, activation of the complement system in the periphery, at the level of nerve injury and DRG, could also contribute to the genesis of neuropathic pain.[31] [32] In the periphery, DF2593A could disrupt the activation and the recruitment of leukocytes, thus attenuating the direct sensitization of the primary nociceptive neurons expressing C5aR.[33] Importantly, DF2593A was effective even when given one week after SNI, and nociceptive hypersensitivity returned to basal values when the DF2593A treatment was halted. These results clearly confirm the therapeutic potential of DF2593A and show that continuous activation of the C5aR is relevant for induction and maintenance of nociceptive hypersensitivity in the SNI model.

In the inflammatory pain models, the actions of C5a appeared to be sequential to PMN activation, which are recruited into the inflammatory site in response to chemokines and lipid mediators.[7] Recent evidence showed that C5aR is expressed in primary nociceptive neurons and that C5a could directly sensitize these fibers.[12] [13] Our studies showed that DF2593A blocked carrageenan and C5a-dependent mechanical hyperalgesia *in vivo* without sedative or central opioid-like effects. Interestingly, DF2593A was also effective in a model of chronic inflammatory pain and in two models of inflammatory arthritic disease even when administered 24 h after the induction of inflammation, sustaining the therapeutic potential of DF2593A and suggesting that C5a is continuously produced and participates in the events leading to maintenance of chronic inflammatory pain. On the other hand, DF2593A did not affect mechanical

hyperalgesia induced by PGE_2 or epinephrine, which directly cause sensitization of primary nociceptive neurons.[6] Therefore, the analgesic effects of DF2593A cannot be ascribed to a non-specific action and is related to its ability to block C5a/C5aR signaling, presumably on both PMN and nociceptive neurons. Finally, it is important to point out that the results obtained with $C5aR^{-/-}$ mice strongly provide target validation/specificity for the effects of DF2593A.

Our findings underline the functional relevance of the minor pocket in GPCR that, though not directly involved in natural ligand binding, cooperates with the fine tuning of the receptor activation process and represents an attractive structural determinant for the rational design of innovative therapeutic compounds. Our pharmacological results not only provide further support to the role of C5a/C5aR signaling in the generation and maintenance of neuropathic pain but also demonstrated that DF2593A represents an innovative non competitive allosteric drug candidate to control pain in multiple therapeutic indications.

4.4 Materials and Methods

4.4.1 Molecular modeling.

The TM domains of CXCR1, CXCR2 and C5aR were identified by sequence alignments with rhodopsin structure by using the MUSCLE software. The C5aR model was refined and compared with CCR5 (pdb code: 4MBS) and further optimized by verifying the GPCR TM-fingerprints alignment.[20] [21] The C5aR TM bundle was assembled and refined as described in the SI and the final C5aR structure was used to dock DF2593A using LiGen.

4.4.2 Cells and migration assay.

PMN and L1.2 cells migration was evaluated using a 48-well micro-chemotaxis chamber as previously described.[19]

4.4.3 Mechanical nociceptive paw test in mice.

Mechanical hyperalgesia was tested in C57BL/6, BALB/C mice and in C5a-receptor-deficient mice ($C5aR^{-/-}$) using the electronic von Frey test as previously reported.[6]

4.4.4 SNI-induced neuropathic pain-like behavior.

The SNI procedure comprised an axotomy and ligation of the tibial and common peroneal nerves leaving the sural nerve intact.[34]

4.4.5 Data analyses and statistics.

For in vivo experiments, results are presented as mean ± SEM. The differences among the groups were compared by ANOVA (one-way) followed by Bonferroni's post hoc test. The level of significance was set at $P<0.05$.

4.5 REFERENCES

1. Kidd, B. L.; Urban, L. A., Mechanisms of inflammatory pain. *British journal of anaesthesia* **2001,** *87* (1), 3-11.
2. Baron, R.; Binder, A.; Wasner, G., Neuropathic pain: diagnosis, pathophysiological mechanisms, and treatment. *The Lancet. Neurology* **2010,** *9* (8), 807-19.

3. Schaffer, D.; Florin, T.; Eagle, C.; Marschner, I.; Singh, G.; Grobler, M.; Fenn, C.; Schou, M.; Curnow, K. M., Risk of serious NSAID-related gastrointestinal events during long-term exposure: a systematic review. *The Medical journal of Australia* **2006,** *185* (9), 501-6.
4. Devulder, J.; Jacobs, A.; Richarz, U.; Wiggett, H., Impact of opioid rescue medication for breakthrough pain on the efficacy and tolerability of long-acting opioids in patients with chronic non-malignant pain. *British journal of anaesthesia* **2009,** *103* (4), 576-85.
5. White, F. A.; Jung, H.; Miller, R. J., Chemokines and the pathophysiology of neuropathic pain. *Proceedings of the National Academy of Sciences of the United States of America* **2007,** *104* (51), 20151-8.
6. Cunha, T. M.; Verri, W. A., Jr.; Silva, J. S.; Poole, S.; Cunha, F. Q.; Ferreira, S. H., A cascade of cytokines mediates mechanical inflammatory hypernociception in mice. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (5), 1755-60.
7. Ting, E.; Guerrero, A. T.; Cunha, T. M.; Verri, W. A., Jr.; Taylor, S. M.; Woodruff, T. M.; Cunha, F. Q.; Ferreira, S. H., Role of complement C5a in mechanical inflammatory hypernociception: potential use of C5a receptor antagonists to control inflammatory pain. *British journal of pharmacology* **2008,** *153* (5), 1043-53.
8. Griffin, R. S.; Costigan, M.; Brenner, G. J.; Ma, C. H.; Scholz, J.; Moss, A.; Allchorne, A. J.; Stahl, G. L.; Woolf, C. J., Complement induction in spinal cord microglia results in anaphylatoxin C5a-mediated pain hypersensitivity. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **2007,** *27* (32), 8699-708.
9. Gerard, C.; Gerard, N. P., C5A anaphylatoxin and its seven transmembrane-segment receptor. *Annual review of immunology* **1994,** *12*, 775-808.
10. Zwirner, J.; Fayyazi, A.; Gotze, O., Expression of the anaphylatoxin C5a receptor in non-myeloid cells. *Molecular immunology* **1999,** *36* (13-14), 877-84.
11. Woodruff, T. M.; Nandakumar, K. S.; Tedesco, F., Inhibiting the C5-C5a receptor axis. *Molecular immunology* **2011,** *48* (14), 1631-42.
12. Jang, J. H.; Clark, J. D.; Li, X.; Yorek, M. S.; Usachev, Y. M.; Brennan, T. J., Nociceptive sensitization by complement C5a and C3a in mouse. *Pain* **2010,** *148* (2), 343-52.
13. Liang, D. Y.; Li, X.; Shi, X.; Sun, Y.; Sahbaie, P.; Li, W. W.; Clark, J. D., The complement component C5a receptor mediates pain and inflammation in a postsurgical pain model. *Pain* **2012,** *153* (2), 366-72.
14. Horuk, R., Chemokine receptor antagonists: overcoming developmental hurdles. *Nature reviews. Drug discovery* **2009,** *8* (1), 23-33.
15. Finch, A. M.; Wong, A. K.; Paczkowski, N. J.; Wadi, S. K.; Craik, D. J.; Fairlie, D. P.; Taylor, S. M., Low-molecular-weight peptidic and cyclic antagonists of the receptor for the complement factor C5a. *Journal of medicinal chemistry* **1999,** *42* (11), 1965-74.
16. Ricklin, D.; Lambris, J. D., Complement-targeted therapeutics. *Nature biotechnology* **2007,** *25* (11), 1265-75.
17. Rosenkilde, M. M.; Benned-Jensen, T.; Frimurer, T. M.; Schwartz, T. W., The minor binding pocket: a major player in 7TM receptor activation. *Trends in pharmacological sciences* **2010,** *31* (12), 567-74.
18. Bertini, R.; Barcelos, L. S.; Beccari, A. R.; Cavalieri, B.; Moriconi, A.; Bizzarri, C.; Di Benedetto, P.; Di Giacinto, C.; Gloaguen, I.; Galliera, E.; Corsi, M. M.; Russo, R. C.; Andrade, S. P.; Cesta, M. C.; Nano, G.; Aramini, A.; Cutrin, J. C.; Locati, M.; Allegretti, M.; Teixeira, M. M., Receptor binding mode and pharmacological characterization of a potent and selective dual CXCR1/CXCR2 non-competitive allosteric inhibitor. *British journal of pharmacology* **2012,** *165* (2), 436-54.
19. Bertini, R.; Allegretti, M.; Bizzarri, C.; Moriconi, A.; Locati, M.; Zampella, G.; Cervellera, M. N.; Di Cioccio, V.; Cesta, M. C.; Galliera, E.; Martinez, F. O.; Di Bitondo, R.; Troiani, G.; Sabbatini, V.; D'Anniballe, G.; Anacardio, R.; Cutrin, J. C.; Cavalieri, B.; Mainiero, F.; Strippoli, R.; Villa, P.; Di Girolamo, M.; Martin, F.; Gentile, M.; Santoni, A.; Corda, D.; Poli, G.; Mantovani, A.; Ghezzi, P.; Colotta, F., Noncompetitive allosteric inhibitors of the inflammatory chemokine receptors CXCR1 and CXCR2: Prevention of reperfusion injury. *Proceedings of the National Academy of Sciences of the United States of America* **2004,** *101* (32), 11791-11796.

20. Fredriksson, R.; Lagerström, M. C.; Lundin, L.-G.; Schiöth, H. B., The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints. *Molecular Pharmacology* **2003,** *63* (6), 1256-1272.
21. Allegretti, M.; Moriconi, A.; Beccari, A. R.; Di Bitondo, R.; Bizzarri, C.; Bertini, R.; Colotta, F., Targeting C5a: recent advances in drug discovery. *Current medicinal chemistry* **2005,** *12* (2), 217-36.
22. Allegretti, M.; Bertini, R.; Bizzarri, C.; Beccari, A.; Mantovani, A.; Locati, M., Allosteric inhibitors of chemoattractant receptors: opportunities and pitfalls. *Trends in pharmacological sciences* **2008,** *29* (6), 280-6.
23. Monk, P. N.; Pease, J. E.; Marland, G.; Barker, M. D., Mutation of aspartate 82 of the human C5a receptor abolishes the secretory response to human C5a in transfected rat basophilic leukemia cells. *European journal of immunology* **1994,** *24* (11), 2922-5.
24. Nikiforovich, G. V.; Baranski, T. J., Structural mechanisms of constitutive activation in the C5a receptors with mutations in the extracellular loops: molecular modeling study. *Proteins* **2012,** *80* (1), 71-80.
25. Moriconi, A.; Cesta, M. C.; Cervellera, M. N.; Aramini, A.; Coniglio, S.; Colagioia, S.; Beccari, A. R.; Bizzarri, C.; Cavicchia, M. R.; Locati, M.; Galliera, E.; Di Benedetto, P.; Vigilante, P.; Bertini, R.; Allegretti, M., Design of noncompetitive interleukin-8 inhibitors acting on CXCR1 and CXCR2. *Journal of medicinal chemistry* **2007,** *50* (17), 3984-4002.
26. Bissantz, C.; Kuhn, B.; Stahl, M., A medicinal chemist's guide to molecular interactions. *Journal of medicinal chemistry* **2010,** *53* (14), 5061-84.
27. Gerber, B. O.; Meng, E. C.; Dötsch, V.; Baranski, T. J.; Bourne, H. R., An Activation Switch in the Ligand Binding Pocket of the C5a Receptor. *Journal of Biological Chemistry* **2001,** *276* (5), 3394-3400.
28. Bizzarri, C.; Pagliei, S.; Brandolini, L.; Mascagni, P.; Caselli, G.; Transidico, P.; Sozzani, S.; Bertini, R., Selective inhibition of interleukin-8-induced neutrophil chemotaxis by ketoprofen isomers. *Biochemical pharmacology* **2001,** *61* (11), 1429-37.
29. Ribeiro, R. A.; Vale, M. L.; Thomazzi, S. M.; Paschoalato, A. B.; Poole, S.; Ferreira, S. H.; Cunha, F. Q., Involvement of resident macrophages and mast cells in the writhing nociceptive response induced by zymosan and acetic acid in mice. *European journal of pharmacology* **2000,** *387* (1), 111-8.
30. Grant, E. P.; Picarella, D.; Burwell, T.; Delaney, T.; Croci, A.; Avitahl, N.; Humbles, A. A.; Gutierrez-Ramos, J. C.; Briskin, M.; Gerard, C.; Coyle, A. J., Essential role for the C5a receptor in regulating the effector phase of synovial infiltration and joint destruction in experimental arthritis. *The Journal of experimental medicine* **2002,** *196* (11), 1461-71.
31. Twining, C. M.; Sloane, E. M.; Schoeniger, D. K.; Milligan, E. D.; Martin, D.; Marsh, H.; Maier, S. F.; Watkins, L. R., Activation of the spinal cord complement cascade might contribute to mechanical allodynia induced by three animal models of spinal sensitization. *The journal of pain : official journal of the American Pain Society* **2005,** *6* (3), 174-83.
32. Li, M.; Peake, P. W.; Charlesworth, J. A.; Tracey, D. J.; Moalem-Taylor, G., Complement activation contributes to leukocyte recruitment and neuropathic pain following peripheral nerve injury in rats. *The European journal of neuroscience* **2007,** *26* (12), 3486-500.
33. Levine, J. D.; Gooding, J.; Donatoni, P.; Borden, L.; Goetzl, E. J., The role of the polymorphonuclear leukocyte in hyperalgesia. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **1985,** *5* (11), 3025-9.
34. Decosterd, I.; Woolf, C. J., Spared nerve injury: an animal model of persistent peripheral neuropathic pain. *Pain* **2000,** *87* (2), 149-58.

Development and Validation Of In Silico Tools For Efficient Library Design and Data Analysis in High Throughput Screening Campaigns

Chapter 5: Molecular Assemblies: Multi-Dimensional Hierarchical Scaffold Analysis

5 MOLECULAR ASSEMBLIES: MULTI-DIMENSIONAL HIERARCHICAL SCAFFOLD ANALYSIS

5.1 INTRODUCTION

The identification of common structural moieties and chemical features present in the large and diverse sets of compounds identified in high-throughput screening (HTS) campaigns is a long standing and unsolved critical point in the hit prioritization phase and in the definition of the chemical follow-up strategy. The establishment of effective preliminary structure activity relationships (SAR) is the first fundamental step that affects the probability of success of the entire lead discovery process. Different approaches have been developed to classify hundreds of thousands of compound hit from HTS campaigns and to search up to millions of discarded compounds looking for false negatives or latent hits.

In our experience those approaches leading to results that can be easily interpreted by a medicinal chemistry perspective are the most effective. The reason is that the final goal of the analysis is to identify the most promising chemical series to be prioritized on the basis of their synthetic feasibility and novelty. Frequently used descriptors often generate good computational models which are, however, not always readily interpretable or simple to explain. Furthermore, they often generate multi chemotype ensembles that do not help answer the final question: “Which chemical series shall I develop?”

A chemical scaffold also referred to as ‘chemotype’ or ‘Markush structure’, can be defined as the common structure characterizing a group of molecules. Molecules sharing the same scaffold are likely to have a similar synthetic pathway as an extension of the concept of molecular template in combinatorial chemistry[1]. Once a scaffold is defined, SARs can be developed by means of the analysis of the effects of the substitution patterns.

Since the first publication of Bemis *et al.*[2], which introduced a systematic analysis of drugs according to their scaffold/framework representation, the interest in the approach based on this representation has increased continuously and it is now a well-established method which flanks molecular descriptors, molecular fingerprints and graphs. Advantages of the approach are: 1) outcomes are both simple to interpret and medicinal chemistry-oriented; 2) some of the most significant features of the graph-based approaches[3] are combined with some others of the molecular fingerprint and the maximum common substructure methods.

The introduction in 2005, by Wilkens *et al.*[4], of hierarchies based on several kinds of scaffold deconstructions represented a milestone in the development of scaffold-based approaches.

In 2007 Schuffenhauer *et al*[5]. demonstrated the potentiality of combining the scaffold-based approach with *ad hoc* graphical representations through the “Scaffold Tree” algorithm and visualization tool; in that work, a rule-based ring disassembly was also introduced. Since then, other decomposition and visualization tools have been developed, such as Scaffold Hunter in 2009[6] and Scaffold Explorer in 2010[7].

In 2008 Gianti and Sartori[8] proposed an alternative procedure to address scaffold decoration, pruning and fragmentation as a workflow for the identification of “privileged fragments”.

Agrofiatis et al.[7], in 2010, demonstrated that the inclusion of relevant side chains and functional groups in the scaffold representation could greatly enhance the derivation of robust SAR, thus indicating that the

explicit consideration of the most significant molecular features overcomes the limits associated with the definition of "*a priori*" pruning rules. In the same year, our group presented the Molecular Assemblies (MA) approach at 18th European Symposium on Quantitative Structure-Activity Relationships (QSAR)[9].

In 2011, Varin et al.[10] proposed an extended version of the previously developed Scaffold Tree and demonstrated that rule-based approaches in fragmentation are less useful and flexible than unbiased ones. Lipkus[11] introduced hierarchies between different abstraction levels and, finally, different hierarchical scaffold decomposition and abstraction approaches were proposed by many authors such as Vogt et al.[12] and Hu et al.[13]

In theory, it is possible to define an arbitrary number of scaffold's representations based on levels of abstraction and pruning rules. Figure 36 shows an example applied to the well-known COX-2 inhibitor celecoxib[14].

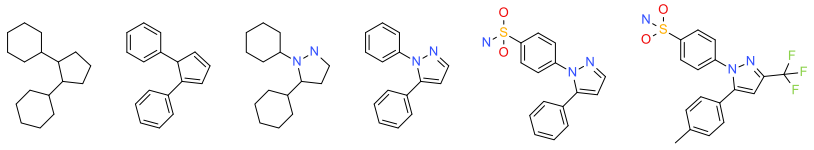


Figure 36. Possible scaffold definitions for celecoxib

A further generalization of Figure 36**a** obtained by removing the ring size and the internal chain length is out of the generally accepted boundaries of what is considered to be a molecular scaffold analysis. It could instead represent the starting point for the molecular graph analysis[15] an alternative, and somehow complementary strategy for data analysis, clustering[16], scaffold hopping[17], and similarity search[18].

Based on our experience of biological data analysis, the main limitation of these implementations (Figure 36**a** to Figure 36**e**) is that a single representation of the scaffold is not sufficient to effectively capture the relationship between the biological activity and the chemical structure of a heterogeneous ensemble of molecules. Actually, most of the methods described so far are based on one unique scaffold abstraction, such as those depicted in Figure 36**c** and Figure 36**d**. These representations are highly effective in congeneric series but quite poor in describing multi scaffold libraries. The problem is that it is not possible to define *a priori* the best representation of a molecule because it depends mainly on the biological context and on the other nearest-neighbors belonging to the screened library. This is particularly true in HTS campaign, in which data analysis and the design of follow-up libraries still lack a general and flexible methodology to derive structural information out of molecular collections and, where possible, to correlate them with real or predicted biological activities. SAR visualization, in other words, still lacks generality (abstraction level) without limiting the need for precise structural information when needed, even if many remarkable approaches have been proposed[19].

In this paper a novel tool named Molecular Assemblies is proposed to overcome restrictions in the use of scaffold analysis based on a single predefined set of rules when diverse libraries are used. We demonstrate here that these limitations affect hit prioritization together with scaffold hopping and SAR expansion in lead optimization settings.

5.2 MOLECULAR ASSEMBLIES.

In HTS campaigns the hits are usually molecules with activity in the micromolar range[20] and in this range of concentration the recognition of the interaction sites are in general ruled by shape and molecular electrostatics more than by more specific and directional interactions. For this reason, it is quite common, for instance, that saturated and aromatic rings show the same activity or that different kinds of chains and functional groups could be exchanged in the molecule while retaining the same level of biological activity. In this scenario, it is not surprising that scaffold representation like Figure 36**a**, Figure 36**b**, and Figure 36**c** perform generally better than the others in the identification of relevant chemotypes, as they are able to cluster together different molecular species which behave similarly in the biological context of the HTS campaign

We intend to demonstrate that representation at a high level of abstraction, like Figure 36**a-c**, is better suited also to cluster molecules characterized at a nanomolar range of activity. As an example, we have interrogated the Integrity™ database searching for COX-2 inhibitors.

Integrity™ (query date: 2012.11.02) collected 1912 Cyclooxygenase-2 Inhibitors, 457 of them in preclinical development or a higher phase. If we use representation Figure 36**a**, as depicted in Figure 37below, a subset of 128 molecules was identified. These 128 molecules correspond to 51 different scaffolds if representation Figure 36**d** (rings and chains connecting rings) is used

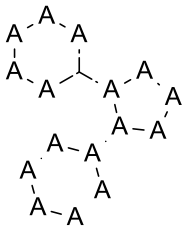


Figure 37. MDL substructure query, of the COX-2 inhibitor common moiety (any atoms and any bonds).

Interestingly, representation Figure 36 includes and clusters together well known marketed drugs such as valdecoxib, celecoxib and carecoxib as well as many others leads and experimental drugs. Whereas in Figure 38. Scaffold representations of 10 of the most diverse cyclooxygenase-2 inhibitors show ten of the most diverse cyclooxygenase-2 scaffold substructures (according to the Tanimoto index calculated by ECFP6 fingerprint) obtained by using representation Figure 36**d**.

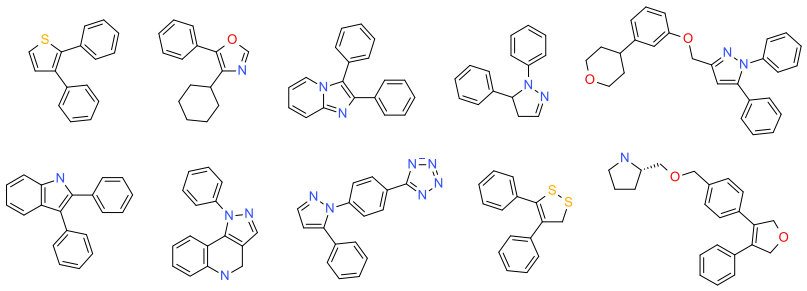


Figure 38. Scaffold representations of 10 of the most diverse cyclooxygenase-2 inhibitors

Inspection of Figure 38, however, suggests another important consideration: 26 out of 51 scaffolds have one or more additional rings; this means that even if the highest level of abstraction is used (Representation Figure 36**a**) more than 50% of the structural scaffolds' information would be lost. If the

entire set is considered, also including molecules in initial biological testing phase, the ratio increases to 130 over 188 scaffolds.

Attempts to extend the structural information coverage has been presented in the past, for example by rule-based decompositions proposed by Schuffenhauer *et al*[21], but, as shown in Figure 39, it is difficult to define a set of rules that maintains a general consistency with the SARs;

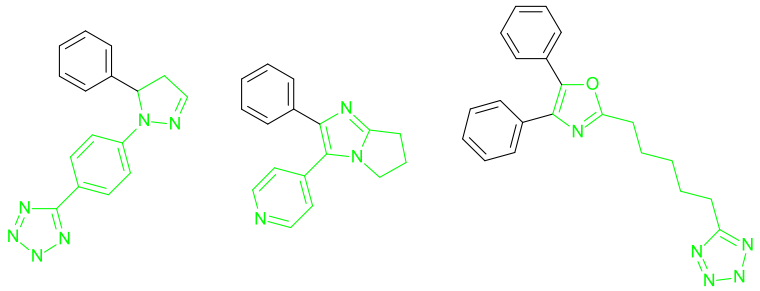


Figure 39. Rule-based scaffold's decompositions are not always able to identify the most useful representation of the molecules. In green a tentative rule-based decomposition is represented; in this case the reasonable hypothesis that phenyl rings are less relevant than heterocyclic ones leads to sub-scaffolds lacking the COX2 moiety.

If representation Figure 36**c** (atom type but not bond type retained) is used and a scaffold-based cluster analysis is performed on the subset of 128 COX-2 inhibitors having only three rings, the outcome is 25 unrelated clusters. In an ideal example of a screening collection of 100,000 molecules, if the 128 active COX-2 inhibitors are identified as hits, the useful information is spread and hidden in more than 500 other unrelated clusters of molecules. On the other hand, if representation Figure 36**a** is used, all the active molecules collapse to a single cluster which probably includes many other inactive molecules. This will be the case especially when scaffolds deconstructions and ring disassembly are implemented; a case in which the enrichment factor associated to the cluster should be low, thus resulting in a de-prioritization with respect to less relevant ones.

It might finally be assumed that representation Figure 36**b** (no atom type but bond order retained) allows a more effective clustering of the important information. In practice, however, it could be demonstrated that the analysis strongly depends on the nature of the dataset, and it is very hard not to bias the analysis if *a priori* suboptimal representation is used.

Thus, we can conclude from this initial example that none of the scaffold representations of Figure 36 is able alone to capture the complexity of the entire ensemble of molecules.

The Drug Discovery Platform at Dompé has started a program to implement cheminformatics tools able to increase the communication between medicinal chemists and cheminformatics to effectively share SAR information and hypotheses, and to plan synthetic strategies for future ad-hoc collections, without the limitations arising from the current technical boundaries of literature methods. As a result of these efforts, we have implemented a tool able to cluster molecules by means of an abstraction method which is more medicinal chemist oriented than those based on a cheminformatics SMARTS rule[22] or fingerprint similarity. This aim has been reached by developing a suite of proprietary tools based on Java Programming language and Pipeline Pilot scripting and this suite, currently employed in our HTS routine data analysis and library design, has been tested by performing a series of experiments hereafter described.

5.3 MATERIALS AND METHODS

5.3.1 Algorithms

Based on the experience accumulated over years of HTS data analysis and design of follow up libraries, we developed the MA toolkit for fast and effective analysis of heterogeneous libraries. Now, we briefly describe the set of rules and the pruning methods so far implemented.

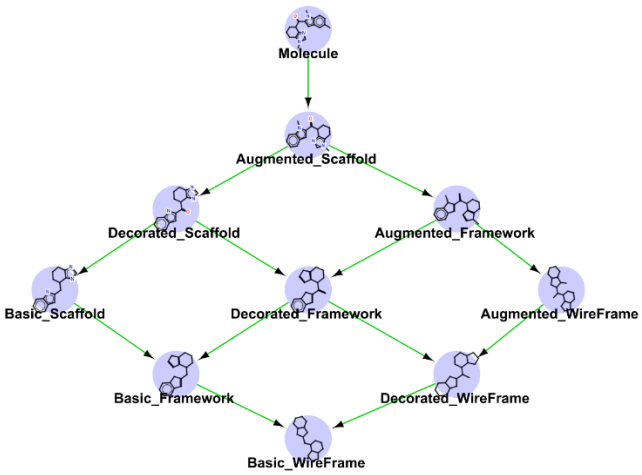


Figure 40. Representation of the nine levels of structure decomposition

As a starting point, we used the widely accepted scaffold abstraction concept (here called *Basic Scaffold*), which, in its simplest representation, is generated by a simple pruning rule, the removal of all side chains and terminal atoms. As a side chain definition is not always straightforward, we have identified a set of nine MA at different abstraction levels, as shown in the Figure 40. Two sets of pruning rules able to define a multidimensional hierarchy have been thus defined. The first rule, which increases chemical abstraction, removes the atom type label (thus generating a Framework) then the bond order (thus generating a Wireframe). The second set of rules is based however, on an increased level of structural information with respect to the basic scaffold: as a first step terminal atoms with bond order greater than one are maintained ("Decorated Scaffold"), and, as second step, the longest atom chain, considering also substitutions, is retained but all terminal non-carbon atoms belonging to side chains are pruned iteratively ("Augmented Scaffold").

One general limitation of scaffold based techniques[4, 6, 21] is that by definition, molecules sharing only partially the same scaffold belongs to distinct clusters.

To overcome this limitation, we have implemented an unbiased fragmentation scheme that can be applied to all nine MA, Figure 41 depicts an example of fragmentation based on an exhaustive and progressive elimination of all the chains from the scaffold.

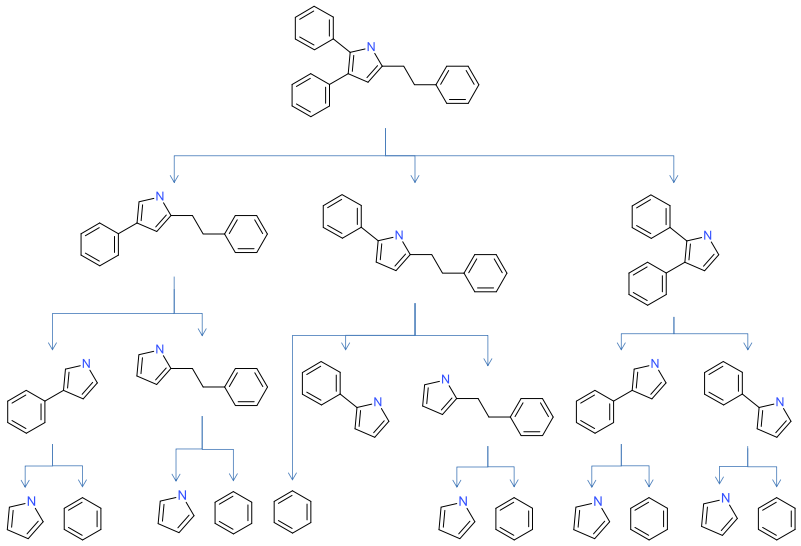


Figure 41. Representation of the hierarchical decomposition of a molecule by the exhaustive and progressive elimination of all the chains from the scaffold.

Unbiased ring disassembly has been also implemented; the methodology for ring decomposition involves the removal of all fused rings, allowing them to open in fragments. The procedure is exemplified in Figure 42.

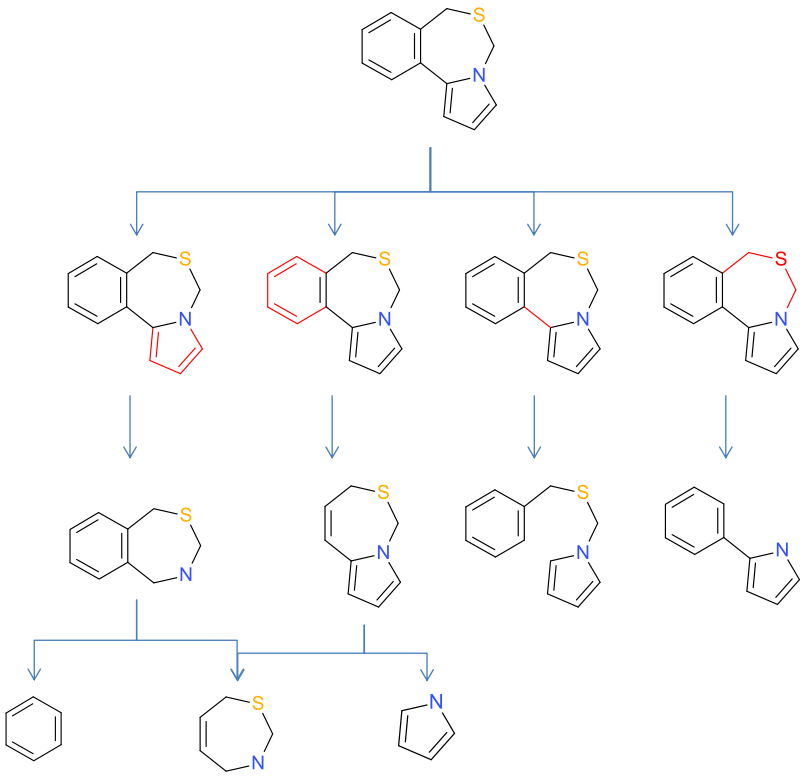


Figure 42. Representation of the hierarchical decomposition of a molecule by the exhaustive and progressive elimination of all fused rings from the scaffold.

For the sake of consistency, we have also introduced a rule to remove internal rings as shown in Figure 43.

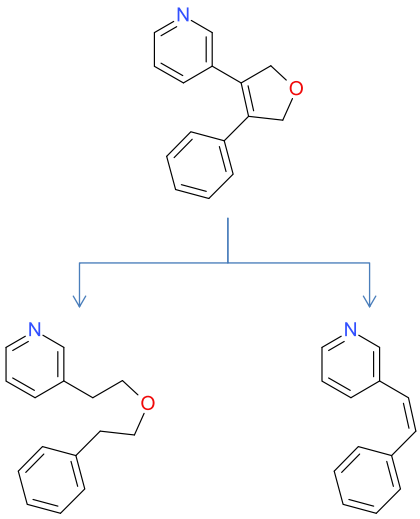


Figure 43. Representation of the hierarchical decomposition of a molecule by the exhaustive and progressive elimination of all internal rings from the scaffold

These fragmentations and deconstructions introduce other hierarchies, meaning that each fragment of the original scaffold is related to all the other representations in a multi-dimensional space.

Because of this multi-dimensional hierarchical scaffold analysis the entire MA produced are highly interconnected and it is possible to move from one to another following SARs. To clarify this concept in Figure 44 is reported an example of how it is possible to combine hierarchies to cluster together molecules with different scaffolds.

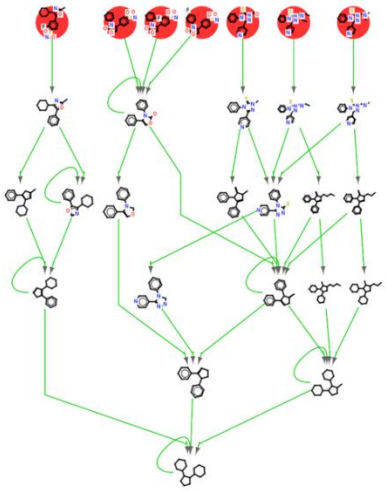


Figure 44.

In the application paragraph a procedure able to generate hierarchies of combinations of representations by the use of network representation is reported

In **Table 5** the output of ring disassembly component is reported; the collected information allows displaying and analyzing results as hierarchies.

Table 5. Components' output of ring disassembly

Frag id	Parent id	N. of frag.	Smiles String	Parent Smiles string	frequency
0	-1	9	C1SCn2cccc2c3ccccc13		1
1	0	9	c1ccc(cc1)c2ccc[nH]2	C1SCn2cccc2c3ccccc13	1
2	0	9	C1SCc2ccccc2C=N1	C1SCn2cccc2c3ccccc13	1
3	2	9	c1ccccc1	C1SCc2ccccc2C=N1	1
4	2	9	C1SCN=CC=C1	C1SCc2ccccc2C=N1	2
5	0	9	C1SCn2cccc2C=C1	C1SCn2cccc2c3ccccc13	1
6	5	9	c1cc[nH]c1	C1SCn2cccc2C=C1	1
7	5	9	C1SCN=CC=C1	C1SCn2cccc2C=C1	2
8	0	9	c1ccc(cc1)CSCn2cccc2	C1SCn2cccc2c3ccccc13	1

The "Frag ID" tag indexes the number of fragments starting from 0 for the parent Molecular Assembly chosen, "Parent frag. Id" establishes the parent-child hierarchy (-1 indicates no parent). The "Number of frag." is the total number of fragments obtained; the SMILES and Parent SMILES strings are respectively the fragment and its parent SMILES strings. The "Frequency" is the occurrence of each fragment in the decomposition.

In our experience, this set of MA and related fragments is able to capture most of the structure activity information coming from HTS campaigns, and it is useful for other applications, as described in the next paragraph. Especially in case of screening of chemically diverse libraries, the combination of fragments, framework and wireframe MA is able to identify relevant chemical moieties in an efficient manner, clustering together many scaffolds with similar shapes despite for example different dispositions of hetero atoms or small differences in bond order. At the moment this is not possible with other known methodologies, like the widely accepted Maximum Common Substructure (MCS).[23] An interesting exploitation of this concept will be found in the paragraph on applications.

5.4 Applications

5.4.1 Compound Clustering & Database Indexing

One of the most critical tasks in the design of large diverse libraries is the mapping of chemical space. Selection of a representative subset of the desired chemical space is generally addressed by the identification of a set of meaningful descriptors[24], a similarity metric for determining the pairwise comparison of molecular structures[25], and a clustering algorithm for grouping structures according to the calculated pairwise similarity values[26]. Many papers compare and analyze clustering algorithms[27], and many clustering techniques are able to address the task for groups of 10^5 to 10^7 compounds. The achievement of chemical consistency within the generated clusters from the point of view of the identification of relevant chemical series is much more difficult. The generation of groupings that can be considered as "series" in a medicinal chemist's perception should be seen as an important asset of the scaffold based techniques. Indeed, the clusters generated by currently available approaches tend to contain an elevated number of scaffolds and this makes it difficult to select chemical series for follow-up activities, although some corrections have been applied to overcome this problem, for example by means of Maximum Overlapping Set (MOS)[28]. For scaffold-based clustering techniques no similarity indices or calculated pairwise similarities are necessary, as the scaffold structure represents itself the similarity index and the aggregation rule, so that each molecule belongs to its own cluster regardless of the nature of the neighbors. In this respect, the approach can be defined as a "Natural Clustering" (NC) approach. Another important feature is that no bias, like the average number of molecules per group or the expected number of scaffolds, has to be introduced. This is very useful when some over-represented scaffolds may drive the analysis. The possibility introduced

by Molecular Assembles to perform hierarchical clustering extends the potentialities of scaffold-based clustering.

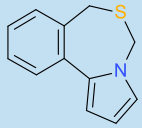
NC makes relational databases ideal for chemical graph-based compound clustering applications. A compound's cluster membership is an objective property as it does not depend on the chemical structure of the other cluster members, and, once the scaffold abstraction is defined, there is no more need to re-cluster the whole library. By organizing molecules in the database by means of clustered SQL indices (based on the MA), it is possible to dramatically reduce the time required for substructure searches, as reported by Wilkens *et al.*[29] and Masciocchi *et al.*[30]. In our implementation, due to the higher abstraction of the frameworks and wireframes, it is also possible to further speed up substructure searches using these representations as a wild character-like approach such as "any atom" or "any bond". Interestingly, MA are *per se* searchable molecular representations, and this allows for the possibility of defining local similarities in substructure searches space. For example, the scaffold of the target molecule could be searched with either a lower or higher similarity to the reference template. Besides that, it is also possible to constrain the local diversity of the scaffold by requiring, for example, a specific hydrogen bond acceptor within the scaffold, or a specific LogP range.

5.4.2 MA as fingerprints

Another intriguing aspect is that it is possible to use the complete ensemble of MA as structural fingerprints[31] with a minimal number of features. This property makes Tanimoto similarity searches[17] very fast and not particularly critical in disk occupation when stored in a database. One possible intrinsic limitation is, however, due to the under estimation of the side chain contribution to the global similarity of the reference molecule with the target ones. On the other hand, our fingerprinting implementation is very useful when it is necessary to identify the most diverse set of compounds, based on a scaffold metrics or when the goal is to find different molecules with a high degree of similarity in terms of scaffold.

To generate the set of fingerprints encoded as signed integers we used the hash function, as implemented in Pipeline Pilot, of the SMILES representation of the MA, as described in **Table 6**.

Table 6. Output properties of the Molecular Assemblies Fingerprint component.

Molecule	Number of fragments	Fragment SMILES	Frag. Freq.	MAFP_DS
	9	C1SCn2cccc2C=C1	1	1809490360
		c1ccccc1	1	-88666468
		C1SCc2ccccc2C=N1	1	227018873
		c1cc[nH]c1	1	471236740
		c1ccc(cc1)c2ccc[nH]2	1	286785889
		C1SCN=CC=C1	2	-2045419586
		C1SCn2cccc2c3ccccc13	1	1411062085
		c1ccc(cc1)CSCn2cccc2	1	-1689663450

In analogy to circular fingerprints[32], where it is possible to use different metrics to generate the fragments[33], it is possible to select scaffolds, frameworks or wireframes as descriptors of the molecule and then to generate a corresponding descriptor's set using fragments and ring disassembly. It is possible to apply the

same procedure (hashing the smiles of the fragment structures as the example reported in figure 2) to our nine representations to calculate MAFP_XX fingerprints from Augmented Scaffold, MAFP_AS, to Basic Wireframe, MAFP_BW.

As a test case, the MA algorithms were applied to a set of 10 well known commercial libraries. All the collections were standardized to hold consistency in protonation states and tautomeric forms (based on "Canonicalization and Enumeration of Tautomers", Sayle and Delany, EuroMUG99, 28-29 October 1999, Cambridge, UK.). Chemical group representations such as nitro groups and salts were removed. Using the standardized structures, internal duplicates in each catalogue were removed, and the Oprea's lead likeness filter was applied[34] thus ending up with a global set of 4.43 million compounds.

Table 7 reports unique structures calculated for each supplier by using the six most relevant representations. From the results it is possible to have a quick view of the chemical space covered by suppliers versus the total number of compounds (complete version of table 3 can be found in the supporting information). This approach makes it very easy to explore scaffold diversity in a library, to compare two libraries, to identify collections with unique or rare scaffolds or to find molecules with the same scaffold or sub scaffold in other collections. Moreover, it is also remarkable that these comparisons take just a few minutes to run more than 4 million compounds.

Table 7. For the most commonly used MA representations of each supplier the number of unique scaffolds are reported.

Vendors	BS	BF	BW	DS	DF	DW
Alinda	17404	7129	2655	18647	10498	6497
Ambinter	323636	97864	24853	356027	173682	92904
Asinex	56104	23953	8581	59964	33660	20945
Enamine	245620	70394	17251	267872	131223	69701
IBS[a]	37160	17011	5772	39310	22908	14727
LC[b]	34143	15279	5378	35387	21321	13840
Princeton[c]	19423	8950	3074	20873	12601	7850
Tim Tec	49896	19755	5821	54371	29297	16538
Uorsy	71034	25934	8959	76280	43495	26529
Vitas-M	61114	25023	7459	65767	36364	21529

BS: basic scaffold, **DS**: decorated scaffold, **BW**: basic wireframe, **DW**: decorated wireframe, **BF**: basic framework, **DF**: decorated framework, a) InterBioScreen, b) Life Chemicals, c) Princeton Biomolecular Research

Another interesting example is the comparison of the chemical similarity among all suppliers by an extended connectivity fingerprints[33b] such as ECFP6 and MA. The results are reported in **Table 8** and **Table 9**. Asymmetric matrices in the two tables account for the percentage of identity in unique ECFP6 fingerprints and unique MAFP_DS, respectively. The latter case is actually the amount of decorated scaffolds (DS) in common with respect to the total number of DS in a library. The matrices are asymmetric because of the different number of DS in each library.

As can be seen directly from color code mapping in **Table 8**, the average similarity (ECFP6 average similarity: 23.4%) is less than that shown in the **Table 9** (MAFP_DS average similarity: 36.6%). It is reasonable to say that scaffold diversity is smaller than molecule diversity. For example, comparing Alinda and Ambinter catalogues according to ECFP6, the first has 9.8% of the diversity of the second catalogue which itself has 14.8% of the diversity of the first one. A fingerprint comparison suggests that the two libraries are mutually diverse, a dramatic change in the results using the DS metrics can be observed:

Ambinter contains 95.9% of the Alinda collection while this latter contains 5% of the Ambinter DSs. Combining the information it is possible to say that the Ambinter contains most of the Alinda scaffolds which, in turn, show a different decoration pattern,. Therefore, depending on the scope of analysis, it is possible to focus on 4.1% of its novel scaffolds or to focus on the common scaffolds in order to increase their diversity. Another interesting comparison: IBS and Life Chemicals (LC) catalogues collect almost the same number of lead-like compounds (IBS: 161317 and LC: 140346) and DSs (IBS: 37160 and LC: 34143). Nevertheless, the two catalogues are very different in both metrics, showing less than 20% of scaffolds and 15% of fingerprints overlap. Therefore, combining the two libraries would enhance the global diversity since they clearly map distinct parts of the chemical space.

Finally, another interesting comparison can be made. Enamine and Uorsy libraries show a fingerprint similarity of around 50% but Uorsy contains 12% of the Enamine DS which in turn contains 41.8% of Uorsy DS. In this case the two libraries are highly complementary in DS. Combined with the high similarity in decoration pattern, this feature can be exploited, for example, in scaffold hopping approaches.

Table 8. Similarity matrix of 10 chemical libraries of screening compounds based in ECFP6 fingerprint.

		ALIDA	AMBINTER	ASINEX	ENAMINE	IBS	LC	PRINCETONBIO	TIMTEC	UORSY	VML
		1	2	3	4	5	6	7	8	9	10
ALIDA	1	100	14.8	18.5	18.5	18.5	18.5	7.4	18.5	11.1	11.1
AMBINTER	2	9.8	100	22	29.3	9.8	17.1	29.3	17.1	41.5	14.6
ASINEX	3	8.3	15	100	16.7	6.7	11.7	5	21.7	8.3	13.3
ENAMINE	4	6.5	15.6	13	100	6.5	13	5.2	15.6	44.2	9.1
IBS	5	8.6	6.9	6.9	8.6	100	10.3	6.9	22.4	6.9	6.9
LC	6	10.6	14.9	14.9	21.3	12.8	100	23.4	19.1	10.6	8.5
PRINCETON	7	5.6	33.3	8.3	11.1	11.1	30.6	100	8.3	33.3	8.3
TIMTEC	8	9.6	13.5	25	23.1	25	17.3	5.8	100	7.7	11.5
UORSY	9	5.1	28.8	8.5	57.6	6.8	8.5	20.3	6.8	100	8.5
VML	10	7.3	14.6	19.5	17.1	9.8	9.8	7.3	14.6	12.2	100

ECFP6

The table reports the percentage of similarity aongst 10 chemical collections of commercially available compounds as from ECFP6 fingerprint analysis. For sake of clarity, a numeric code has been added to each supplier's name. The color code ranges from red (0%) to green (100%).

Table 9. Similarity matrix of 10 chemical libraries of screening compounds based in DS fingerprint

		ALIDA	AMBINTER	ASINEX	ENAMINE	IBS	LC	PRINCETONBIO	TIMTEC	UORSY	VML
		1	2	3	4	5	6	7	8	9	10
ALIDA	1	100	95.9	36.4	43.1	28.9	16.5	35.7	82	24.9	87.1
AMBINTER	2	5	100	7.6	47.3	10.1	8	5.6	13.9	7.9	17.2
ASINEX	3	11.3	45	100	19	16.7	8.7	15.7	27.2	8.9	27.1
ENAMINE	4	3	62.8	4.2	100	3.8	3.3	3	6.1	11.9	6.6
IBS	5	13.7	91.4	25.4	25.7	100	14.8	23.2	39.9	11.1	83.2
LC	6	8.7	80.8	14.7	25.1	16.5	100	15.3	25.1	10.6	24.9
PRINCETON	7	31.9	95.9	45.1	39	43.7	26	100	73.9	19.1	85.7
TIMTEC	8	28.1	91.3	30	29.9	28.8	16.3	28.4	100	13.8	67.3
UORSY	9	6.1	36.7	7	41.8	5.7	4.9	5.2	9.8	100	10.8
VML	10	24.7	92.9	24.7	26.9	49.7	13.4	27.2	55.7	12.5	100

Decorated Scaffold Inchikey

The table contains the percentage of the similarity among a selection of chemical collections of commercially available compounds using DS. A numeric code has been added to each supplier's name for clarity. The color code ranges from red (0%) to green (100%).

If large chemical collections have to be compared, the proposed approach can be considered complementary to others reported in literature[35]. To highlight differences between the two descriptors, the principal component analysis of the two similarity matrices is reported in Figure 45.

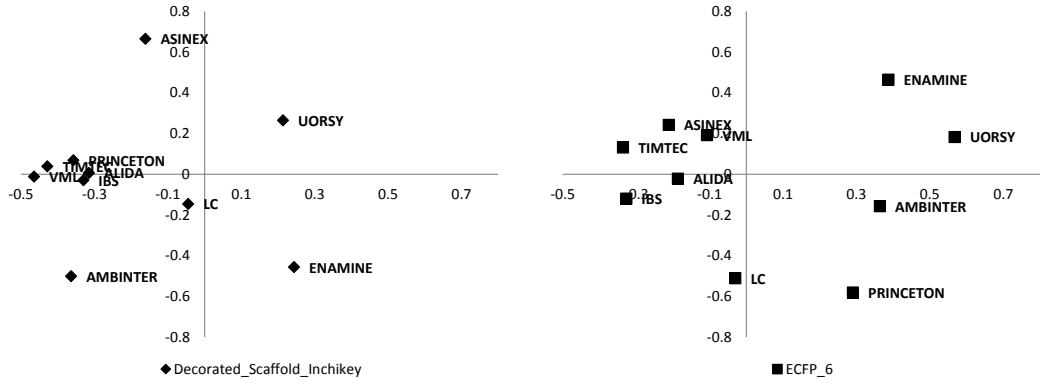


Figure 45.

5.4.3 Molecular Assembly based library design

Many authors have addressed the challenge of designing both combinatorial[36] and diversity-driven libraries for specific[37] and generic applications[38] by focusing on reagents and products.[39],[37a, 40]

With the aim of designing a library driven by scaffold diversity or follow-up libraries for SAR expansion (as from primary HTS data), we propose here two approaches based on the selection of commercially available compounds[41].

Select commercially available compounds seems to be very useful for many reasons, the most relevant being: 1- using stock compounds it is possible to rapidly evaluate SAR reducing costs and delivery time[41]; 2- it is not necessary to limit, at an early stage, the chemistry to a few chemotypes based on the primary hits scaffolds; 3- as will be described in the next paragraph, the multi-scaffold based procedure allows the control of chemotype diversity without dispersion because a significant number of analogues for each scaffold is retrieved. Thus the library is made of a collection of diverse chemotypes miniseries. A good example of chemical space mapping, according to a hierarchical scaffold metric, is reported by Langdon *et al.*[42]. where a scaffold tree[5] decomposition in combination with Tree Maps[43] as visualization tool is shown to be an effective way of illustrating both the distribution and structural diversity of chemical scaffolds.

It is possible to apply MA to the design of follow-up focused libraries for discovery projects. The procedure is based on the ability of hierarchies to constrain the shape of molecules and, at the same time, to fine tune the diversity and the properties of scaffolds.

In Figure 46. a schematic representation of this approach is depicted. Once a wireframe is selected, the library is created on the derived tree in the basic wireframe > framework > scaffold dimension.

The core procedure is implemented in our protocol as follows:

Ring 1: Select a parent wireframe of interest;

Ring 2: Define a range of values for derived frameworks that have to be represented in the library and define the objective function (for example maximize diversity);

Ring 3: Define a range of values for derived scaffolds, as per selection in the previous step, and define the objective function (for example maximize diversity);

Ring 4: Define a range of values of molecules derived from each scaffold selected in the previous step and define the objective functions for each scaffold library and for the global one (for example maximize diversity, drug likeness …)

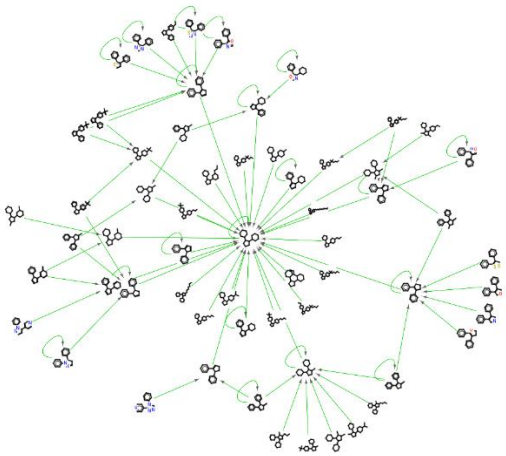


Figure 46.

The same procedure used to explore a single wireframe could be expanded to an arbitrary number of such. The wireframes can have the greatest diversity or be related by fragmentation or ring disassembly rules. The latter approach could be successful when used to expand corporate libraries for singletons removal strategies and to populate new chemical space. Methodologies to specifically address diversity in the scaffold space have already been described[44]. We found useful the proposed MAFP, as described in the previous section.

The balance between diversity, to achieve a good coverage of the desired chemical space, and the possibility to embed information useful for SAR analysis, is a common issue in the design of a general library. Clustering based on similarity index using properties or fingerprint descriptors is very useful in the selection of diverse molecules but, depending on the algorithm, it can be difficult to obtain chemotype-oriented clusters.

The concept of chemotype is, of course, questionable but in general it is expected to carry out a set of hits able to generate a series that could become the prototype of the Markush structure.[45] In turn, the Markush structure can be considered conceptually as a scaffold extension (for us a molecular assembly). The reason of the frequent inconsistency between clustering in terms of scaffolds is mainly based on marked differences in scaffold populations rather than in similarity indexes or descriptors. When the algorithm has to cluster thousands of singletons and thousands of molecules belonging to hundreds or even thousands of scaffolds, it tends in general to split highly populated scaffolds into different clusters and to aggregate the less populated ones. In the reference dataset (4.4 million lead-like compounds from 10 vendors) 70% of all BS have less than 4 analogues and 34% of all BS are singletons. At the same time the most populated BS (phenyl ring) have 228752 analogues, and 338 BS having more than 1000 analogues show a mean of 3503 compounds per cluster (standard deviation is 12752), accounting for 1180769 molecules, that is the 27% of the entire set. Summarizing, 61% of the available chemical space is composed by singletons or by overpopulated scaffolds; in this scenario, a scaffold normalization technique is needed to retain chemotype representation in the library design procedure.

This problem could be minimized by using the Jarvis-Patrick (JP) clustering method. This method is applicable up to medium sized sets (up to 10^5 compounds) because of the $O(N^2)$ dependency in time and O(N) in space (where N is the number of items to cluster); as a consequence, a substantial amount of time is required if N is large. In other words, if a single processor is used to generate the NxN correlation matrix then more than 30 days of calculations are needed to analyze 1 million compounds. Hierarchical

(agglomerative) methods show in general a $O(N^3)$ dependency in time and $O(N^2)$ in space; it could be reduced, in the reciprocal nearest neighbor (RNN) version of Ward's method[46] to the same dependency as in JP method.

Using fast clustering methods like MAXMIN[47] method of Waldman O(nN) as the one available in Pipeline Pilot[48] it is possible to handle millions of compounds but it is generally hard to control the sparseness of the scaffolds, and some bias like the average number of molecules per scaffold must be introduced. In this paper we propose to use methods based on the hierarchy of the MA to generate diverse scaffold-constrained libraries embedding expected scaffold populations from the beginning with a dependency in time of O(N) that may be calculated only once. We calculate the MA during the compounds registration procedure without the need for recalculation.

5.4.4 Network analysis

One of the most promising applications of MA is in the analysis of large datasets from HTS. Many authors proposed effective tools for HTS data analysis.[49,50,51,52] Recently, Scaffold Trees have been successfully applied to this kind of data[5].

As previously demonstrated by the COX2 inhibitors case example, the higher flexibility of MA is able to capture most of the useful information; it is indeed possible to derive trees in multiple dimensions such as wireframe > framework > scaffold or augmented > decorated > basic or wireframe > ring disassembly > fragments, and all other possible directions maximizing the information of the dataset. The hierarchical representation is very useful, but it does not allow as full a graphical representation of dataset information as it could instead being extracted by means of MA in a single visualization.

To overcome this problem we propose a network representation of the dataset. The MA approach has been combined with the network representation in order to generate a more convenient tool for SAR evaluation and visualization.[53,54,55,56]

This procedure is able to generate a convenient representation of the dataset according to the SAR, and it focuses on the assemblies which are relevant to their generation. This reduced representation still encodes the hierarchy of the MA, this feature conveniently guides the user from one molecular assembly to another in the direction of increasing or decreasing levels of abstraction.

A procedure aimed at clustering and representing them as a network is now proposed and the core procedure is the following:

- Cluster compounds based on the MA
- Weighing them by the enrichment factor
- Generating a connection table for each cluster having at least a molecule in common with another one.
- Weighing the connections according to the number of molecules shared by the clusters.

Once the data matrix has been generated, it is possible to visualize it by means of many different software packages[54]. Filtering by enrichment factors and by connection weights makes it possible to focus on the most relevant information. In Figure 47. a cartoon showing a possible outcome of this analysis is depicted.

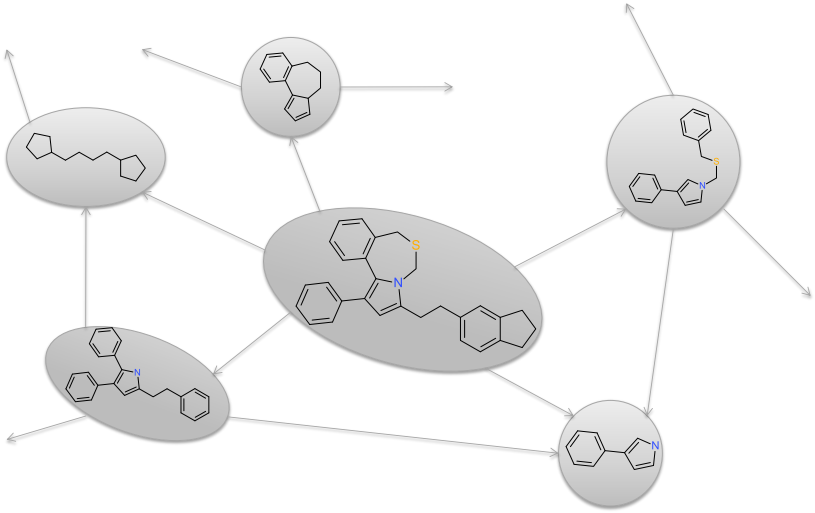


Figure 47.

As evidenced by Figure 47. the graphical representation of the data matrix consists of an oriented network, where nodes are in general molecular fragments or entire small molecules, and the direction of the edges defined by the fragmentation rule starts from the originator fragment and ends up with the corresponding fragments. In general, the obtained network is always fully connected, because small fragments corresponding for example to the benzene ring are shared by a large group of the original molecules contained in the fragmented library. While a quantitative treatment of the obtained network will be discussed elsewhere, we anticipate here some qualitative considerations. As a first point, it is reasonable that highly connected singletons tend to be small fragments shared by a large number of molecules included in the library, whereas low molecular weight singletons, possessing low connectivity, represent potential interesting decorations of some specific group of the originator molecules. If this particular group is enriched in a specific activity of interest, then the low molecular weight singleton connecting all the molecules included in the group could be a pharmacophore.

As a second point, high molecular weight singletons derived from the fragmentation procedure, which connect a cluster of molecules with enriched activity, could represent chemical scaffolds or the “minimal chemical entity” conferring the selected activity to the connected cluster.

As a third point, it is understandable that the significance of the singleton composing the networks may change according to the fragmentation rules. While the approach suggested herein consists of a purely informatics fragmentation procedure, an alternative method is conceivable, where a network is formed only by singletons consisting in reaction intermediates derived from chemical procedures used to produce the molecules included in the original library. In other words, the “fragments” included in such a network would be the precursors used to synthesize larger molecules. If this is the case, then the pathways connecting any two singletons also represent a possible synthetic strategy to be used to attach a specific interesting low molecular weight singleton to another representing, for example, a scaffold.

In our experience connections between chemotypes are not easily identified without the molecular assembly’s decomposition. Filtering by enrichment factor and ranking by number of connections of each cluster make it possible to focus on the highly connected singletons. These molecules are important because they connect different chemotypes without overlapping fragments, giving a better chance to find the most relevant parts of active molecules, the fragments that could be exchanged, the bond order and the atom type that are relevant for SAR derivation. This approach allows us to start the SAR analysis from

the molecules which are generally underestimated singletons and molecules with small ligand efficacy, because these are generally connected to relevant clusters of compound series; in this scenario the additional valuable information could be embedded in the SAR of the major hit series to extend and connect different series one to another. This approach could be considered an extension of the already proposed compound set enrichment[10, 57], by implementing a higher level of abstraction it is potentially able to identify additional latent hit series[58] which are connected with conventional hit series.

5.5 CONCLUSIONS AND FUTURE DIRECTIONS

We propose MA as a fast and flexible method for the analysis of the chemical space, library design and SAR studies.

This set of tools could be useful in the management of large compound collections as in the analysis of HTS campaign results as well as in focused libraries design. On the other hand, by this kind of data organization, the analysis of scaffold-activity relationships is very efficient in the identification of relevant clusters and in the connection of different chemotypes by means of biological activity. The limitation in the underestimation of the side chain effect can be easily overcome by the joined concurrent of other techniques such as molecular matched pairs[59] (MMP). In this case the already identified cores (the MA) makes MMP equally as or even more efficient[60] or consistent[61] than other methods. Furthermore, using MA with a higher level of abstraction, it is possible to compare effectively SAR behaviors on multiple scaffolds, or support scaffold hopping strategies.

Another interesting application, still in an early phase of evaluation, is the annotation of the library according to the categorized drugs by therapeutic areas with the aim of accelerating the identification of target-based or disease-based libraries using for example annotated database such as MDDR, CMC or WOMBAT[62], or public databases[63].

An evolution of MA could be the possibility of defining a set of rules at runtime. It allows the selection of the most suitable MA without any previous knowledge or bias. As previously discussed, the definition of a fixed set of rules introduces a bias that could be minimized, not completely removed, using our set of nine molecular representations. By the introduction of the Generalized Molecular Assemblies (GMA) engine, it is possible to change the set of rules iteratively, with many useful procedures like genetic algorithms, stochastic approaches or, in some cases, also an exhaustive enumeration of all possible combinations, once the optimization function is defined.

By this approach it is possible to identify the molecular representation able to extract the most relevant information contained in the data sets.

5.6 REFERENCES

1. Katritzky, A. R.; Kiely, J. S.; Hebert, N.; Chassaing, C., Definition of Templates within Combinatorial Libraries. *J Comb Chem* **2000,** *2* (1), 2-5.
2. Bemis, G. W.; Murcko, M. A., The properties of known drugs. 1. Molecular frameworks. *J Med Chem* **1996,** *39* (15), 2887-93.
3. Bonchev, D.; Rouvray, D. H., *Chemical graph theory : introduction and fundamentals*. Abacus: New York ; London, 1991.

4. Wilkens, S. J.; Janes, J.; Su, A. I., HierS: hierarchical scaffold clustering using topological chemical graphs. *J Med Chem* **2005,** *48* (9), 3182-93.
5. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H., The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. *J Chem Inf Model* **2007,** *47* (1), 47-58.
6. Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H., Interactive exploration of chemical space with Scaffold Hunter. *Nat Chem Biol* **2009,** *5* (8), 581-3.
7. Agrafiotis, D. K.; Wiener, J. J., Scaffold explorer: an interactive tool for organizing and mining structure-activity data spanning multiple chemotypes. *J Med Chem* **2010,** *53* (13), 5002-11.
8. Gianti, E.; Sartori, L., Identification and selection of "privileged fragments" suitable for primary screening. *J Chem Inf Model* **2008,** *48* (11), 2129-39.
9. Tsantili-Kakoulidou, A.; Agrafiotis, D. K., The 18th European symposium on quantitative structure-activity relationships. *Expert Opin Drug Discov* **2011,** *6* (4), 453-6.
10. Varin, T.; Schuffenhauer, A.; Ertl, P.; Renner, S., Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. *J Chem Inf Model* **2011,** *51* (7), 1528-38.
11. Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F., 3rd; Schenck, R. J.; Trippe, A. J., Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J Org Chem* **2008,** *73* (12), 4443-51.
12. Vogt, M.; Huang, Y.; Bajorath, J., From activity cliffs to activity ridges: informative data structures for SAR analysis. *J Chem Inf Model* **2011,** *51* (8), 1848-56.
13. Hu, Y.; Stumpfe, D.; Bajorath, J., Lessons learned from molecular scaffold analysis. *J Chem Inf Model* **2011,** *51* (8), 1742-53.
14. Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; AndersonGd; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y. Y.; Isakson, P. C., Synthesis and biological evaluation of the 1,5-diarylpyrazole class of cyclooxygenase-2 inhibitors: identification of 4-[5-(4-methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]benze nesulfonamide (SC-58635, celecoxib). *J Med Chem* **1997,** *40* (9), 1347-65.
15. (a) Amigo, J. M.; Galvez, J.; Villar, V. M., A review on molecular topology: applying graph theory to drug discovery and design. *Naturwissenschaften* **2009,** *96* (7), 749-61; (b) Grave, K. D.; Costa, F., Molecular graph augmentation with rings and functional groups. *J Chem Inf Model* **2010,** *50* (9), 1660-8.
16. Gardiner, E. J.; Gillet, V. J.; Willett, P.; Cosgrove, D. A., Representing clusters using a maximum common edge substructure algorithm applied to reduced graphs and molecular graphs. *J Chem Inf Model* **2007,** *47* (2), 354-66.
17. Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J., Scaffold hopping using clique detection applied to reduced graphs. *J Chem Inf Model* **2006,** *46* (2), 503-11.
18. Gillet, V. J.; Willett, P.; Bradshaw, J., Similarity searching using reduced graphs. *J Chem Inf Comput Sci* **2003,** *43* (2), 338-45.
19. (a) Kolpak, J.; Connolly, P. J.; Lobanov, V. S.; Agrafiotis, D. K., Enhanced SAR maps: expanding the data rendering capabilities of a popular medicinal chemistry tool. *J Chem Inf Model* **2009,** *49* (10), 2221-30; (b) Prasanna, M. D.; Vondrasek, J.; Wlodawer, A.; Rodriguez, H.; Bhat, T. N., Chemical compound navigator: a web-based chem-BLAST, chemical taxonomy-based search engine for browsing compounds. *Proteins* **2006,** *63* (4), 907-17; (c) Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J., Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discov Today* **2010,** *15* (15-16), 630-9.
20. Mestres, J.; Veeneman, G. H., Identification of "latent hits" in compound screening collections. *J Med Chem* **2003,** *46* (16), 3441-4.
21. Ertl, P.; Schuffenhauer, A.; Renner, S., The scaffold tree: an efficient navigation in the scaffold universe. *Methods Mol Biol* **2011,** *672*, 245-60.

22. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* **1988,** *28*, 31-36.
23. Hariharan, R.; Janakiraman, A.; Nilakantan, R.; Singh, B.; Varghese, S.; Landrum, G.; Schuffenhauer, A., MultiMCS: a fast algorithm for the maximum common substructure problem on multiple molecules. *J Chem Inf Model* **2011,** *51* (4), 788-806.
24. Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C.; Glick, M.; Davies, J. W., How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J Chem Inf Model* **2009,** *49* (1), 108-19.
25. Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P., Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J Chem Inf Model* **2012**.
26. (a) Brown, R. D.; Martin, Y. C., Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J Chem Inf Comput Sci* **1996,** *36* (3), 572-584; (b) McGregor, M. J.; Pallai, P. V., Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J Chem Inf Comput Sci* **1997,** *37* (3), 443-448.
27. Raymond, J. W.; Blankley, C. J.; Willett, P., Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. *J Mol Graph Model* **2003,** *21* (5), 421-33.
28. Stahl, M.; Mauser, H.; Tsui, M.; Taylor, N. R., A robust clustering method for chemical structures. *J Med Chem* **2005,** *48* (13), 4358-66.
29. Wilkens, S. J., Relational database driven two-dimensional chemical graph analysis. *Chem Biol Drug Des* **2006,** *68* (3), 135-8.
30. Masciocchi, J.; Frau, G.; Fanton, M.; Sturlese, M.; Floris, M.; Pireddu, L.; Palla, P.; Cedrati, F.; Rodriguez-Tome, P.; Moro, S., MMsINC: a large-scale chemoinformatics database. *Nucleic Acids Res* **2009,** *37* (Database issue), D284-90.
31. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A., Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J Chem Inf Comput Sci* **2004,** *44* (3), 1177-85.
32. Glem, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J., Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006,** *9* (3), 199-204.
33. (a) Godden, J. W.; Stahura, F. L.; Bajorath, J., Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. *J Chem Inf Model* **2005,** *45* (6), 1812-9; (b) Rogers, D.; Hahn, M., Extended-connectivity fingerprints. *J Chem Inf Model* **2010,** *50* (5), 742-54.
34. Hann, M. M.; Oprea, T. I., Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* **2004,** *8* (3), 255-63.
35. (a) Kogej, T.; Blomberg, N.; Greasley, P. J.; Mundt, S.; Vainio, M. J.; Schamberger, J.; Schmidt, G.; Huser, J., Big pharma screening collections: more of the same or unique libraries? The AstraZeneca-Bayer Pharma AG case. *Drug Discov Today* **2012**; (b) Engels, M. F.; Gibbs, A. C.; Jaeger, E. P.; Verbinnen, D.; Lobanov, V. S.; Agrafiotis, D. K., A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. *J Chem Inf Model* **2006,** *46* (6), 2651-60; (c) Schamberger, J.; Grimm, M.; Steinmeyer, A.; Hillisch, A., Rendezvous in chemical space? Comparing the small molecule compound libraries of Bayer and Schering. *Drug Discov Today* **2011,** *16* (13-14), 636-41.
36. Nilakantan, R.; Nunn, D. S., A fresh look at pharmaceutical screening library design. *Drug Discov Today* **2003,** *8* (15), 668-72.
37. (a) Stocks, M. J.; Wilden, G. R.; Pairaudeau, G.; Perry, M. W.; Steele, J.; Stonehouse, J. P., A practical method for targeted library design balancing lead-like properties with diversity. *ChemMedChem* **2009,** *4* (5), 800-8; (b) Balakin, K. V.; Ivanenkov, Y. A.; Savchuk, N. P., Compound library design for target families. *Methods Mol Biol* **2009,** *575*, 21-46; (c) Herdemann, M.; Heit, I.; Bosch, F. U.; Quintini, G.; Scheipers, C.; Weber, A., Identification of potent ITK inhibitors through focused compound library design including structural information. *Bioorg Med Chem Lett* **2010,** *20* (23), 6998-7003; (d) Harris, C. J.; Hill, R. D.; Sheppard, D. W.; Slater, M. J.; Stouten, P. F., The design and application of target-focused compound

libraries. *Comb Chem High Throughput Screen* **2011,** *14* (6), 521-31; (e) Gregori-Puigjane, E.; Mestres, J., Coverage and bias in chemical library design. *Curr Opin Chem Biol* **2008,** *12* (3), 359-65.
38. (a) Yeap, S. K.; Walley, R. J.; Snarey, M.; van Hoorn, W. P.; Mason, J. S., Designing compound subsets: comparison of random and rational approaches using statistical simulation. *J Chem Inf Model* **2007,** *47* (6), 2149-58; (b) Gorse, A. D., Diversity in medicinal chemistry space. *Curr Top Med Chem* **2006,** *6* (1), 3-18; (c) Gillet, V. J., New directions in library design and analysis. *Curr Opin Chem Biol* **2008,** *12* (3), 372-8.
39. Jhoti, H., Fragment-based drug discovery using rational design. *Ernst Schering Found Symp Proc* **2007,** (3), 169-85.
40. Akella, L. B.; DeCaprio, D., Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr Opin Chem Biol* **2010,** *14* (3), 325-30.
41. (a) Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A., Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J Chem Inf Model* **2010,** *50* (4), 470-9; (b) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E., Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J Chem Inf Comput Sci* **2004,** *44* (2), 643-51.
42. Langdon, S. R.; Brown, N.; Blagg, J., Scaffold diversity of exemplified medicinal chemistry space. *J Chem Inf Model* **2011,** *51* (9), 2174-85.
43. Kibbey, C.; Calvet, A., Molecular Property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. *J Chem Inf Model* **2005,** *45* (2), 523-32.
44. Li, R.; Stumpfe, D.; Vogt, M.; Geppert, H.; Bajorath, J., Development of a method to consistently quantify the structural distance between scaffolds and to assess scaffold hopping potential. *J Chem Inf Model* **2011,** *51* (10), 2507-14.
45. Cosgrove, D. A.; Green, K. M.; Leach, A. G.; Poirrette, A.; Winter, J., A system for encoding and searching Markush structures. *J Chem Inf Model* **2012,** *52* (8), 1936-47.
46. Downs, G. M.; Barnard, J. M., Clustering Methods and Their Uses in Computational Chemistry. In *Reviews in Computational Chemistry*, John Wiley & Sons, Inc.: 2003; pp 1-40.
47. Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M., Optimization and visualization of molecular diversity of combinatorial libraries. *Mol Divers* **1996,** *2* (1-2), 64-74.
48. Accelrys *Pipeline Pilot*, 8.0.1; San Diego, 2010.
49. Bocker, A.; Schneider, G.; Teckentrup, A., NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. *J Chem Inf Model* **2006,** *46* (6), 2220-9.
50. Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J., SARANEA: a freely available program to mine structure-activity and structure-selectivity relationship information in compound data sets. *J Chem Inf Model* **2010,** *50* (1), 68-78.
51. Richon, A., LeadScope: data visualization for large volumes of chemical and biological screening data. *J Mol Graph Model* **2000,** *18* (1), 76-9.
52. Nicolaou, C. A.; Apostolakis, J.; Pattichis, C. S., De novo drug design using multiobjective evolutionary graphs. *J Chem Inf Model* **2009,** *49* (2), 295-307.
53. Xiong, B.; Liu, K.; Wu, J.; Burk, D. L.; Jiang, H.; Shen, J., DrugViz: a Cytoscape plugin for visualizing and analyzing small molecule drugs in biological networks. *Bioinformatics* **2008,** *24* (18), 2117-8.
54. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **2003,** *13* (11), 2498-504.
55. Iyer, P.; Stumpfe, D.; Bajorath, J., Molecular mechanism-based network-like similarity graphs reveal relationships between different types of receptor ligands and structural changes that determine agonistic, inverse-agonistic, and antagonistic effects. *J Chem Inf Model* **2011,** *51* (6), 1281-6.
56. Lepp, Z.; Huang, C.; Okada, T., Finding key members in compound libraries by analyzing networks of molecules assembled by structural similarity. *J Chem Inf Model* **2009,** *49* (11), 2429-43.

57. Varin, T.; Gubler, H.; Parker, C. N.; Zhang, J. H.; Raman, P.; Ertl, P.; Schuffenhauer, A., Compound set enrichment: a novel approach to analysis of primary HTS data. *J Chem Inf Model* **2010,** *50* (12), 2067-78.
58. Varin, T.; Didiot, M. C.; Parker, C. N.; Schuffenhauer, A., Latent hit series hidden in high-throughput screening data. *J Med Chem* **2012,** *55* (3), 1161-70.
59. (a) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J., Matched molecular pairs as a medicinal chemistry tool. *J Med Chem* **2011,** *54* (22), 7739-50; (b) Wassermann, A. M.; Bajorath, J., Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Med Chem* **2011,** *3* (4), 425-36; (c) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B., Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J Med Chem* **2006,** *49* (23), 6672-82.
60. Hussain, J.; Rea, C., Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J Chem Inf Model* **2010,** *50* (3), 339-48.
61. Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J., MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. *J Chem Inf Model* **2012,** *52* (5), 1138-45.
62. Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L., Predicting new molecular targets for known drugs. *Nature* **2009,** *462* (7270), 175-81.
63. Zhou, Y.; Zhou, B.; Chen, K.; Yan, S. F.; King, F. J.; Jiang, S.; Winzeler, E. A., Large-scale annotation of small-molecule libraries using public databases. *J Chem Inf Model* **2007,** *47* (4), 1386-94.

Development and Validation Of In Silico Tools For Efficient Library Design and Data Analysis in High Throughput Screening Campaigns

Chapter 6: Molecular Assemblies For HTS Chemical Library Design

6 MOLECULAR ASSEMBLIES FOR HTS CHEMICAL LIBRARY DESIGN

6.1 INTRODUCTION

A sizeable and properly organized chemical library is a fundamental prerequisite for any High Throughput Screening (HTS) project. The quality of a chemical library significantly affects the rate of success of HTS campaigns.

A sizeable and properly organized chemical library is a fundamental prerequisite for any HTS project. The quality of a chemical library significantly influences the rate of success of HTS campaigns[1,2].The creation of a high-quality chemical library is therefore of paramount importance in increasing the odds of finding active compounds through HTS. Thus, the creation of new schemes to build structured and diverse compound collections is an active research field that continuously evolves in tandem with the needs of increasing accuracy and specificity while reducing the total number of compounds to be screened, in order to minimize false positive and false negative[3].

The most critical aspects of creating a chemical collection are drug-likeness and analytical properties (i.e., identity and purity) of the selected molecules. The number of compounds present in commercial collections violating Lipinski rules[4] or carrying reactive groups responsible to produce interference with biochemical assays is considerably high[5].

Further, other important issues of any chemical library are chemical diversity and chemotype representativeness, which are two features that inevitably compete with each another, especially if the chemical collection is of limited size. Even when all these characteristics are well balanced within a diverse chemical library, however the chances of false positive results from HTS hits remain relatively high, motivating the search for more effective strategies to perform HTS campaigns with higher efficiency.

To build a chemical collection of approximately 200,000 high quality commercial molecules was the aim of the Italian Drug Discovery Network (IDDN) initiative, whose participants are Angelini, Dompé, IEO – European Institute of Oncology, IIT – Italian Institute of Technology, Recordati, Rottapharm-Madaus. Toward this aim, we first selected a collection of drug-likeness compounds, according to standard criteria reported hereafter. We then pursued the objective of creating a diverse compound collection able to return a preliminary structure-activity relationship (SAR) from possible HTS hits. This challenge was motivated by the central idea that a preliminary SAR from HTS hits could greatly strengthen the hit identification process, reducing the rate of false positives and false negatives via the recovery of close analogues from congeneric series.

A significant effort was therefore first devoted to scaffold analysis for the creation of a high quality and diverse IDDN chemical library. As a means to optimize scaffold selection and clustering, we here introduce the concept of super- and sub- scaffold. Briefly, given any scaffold according to the Molecular Assemblies definition, a sub-scaffold is part of the parent scaffold lacking at least one ring system, while a super-scaffold has at least one ring system more than the parent one (Figure 48). Super- and sub- scaffolds in this contest, are the subset of all possible fragments of the Decorated Scaffold representation that correspond to the decorated scaffold of another molecule (Figure 50).

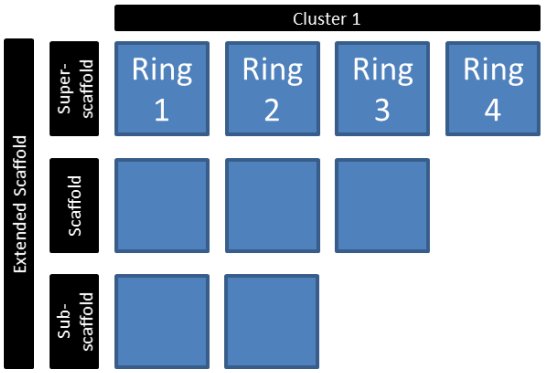


Figure 48. Schematic Representation of the definition of super-, sub-scaffold and extended scaffold

The IDDN collection included a minimal number of representatives for each of the selected chemotype (scaffold-based). Although limiting chemical diversity, this was an essential step toward the identification of preliminary SARs centered on potential HTS hits. Considering that in many cases the first HTS campaign is performed with a single replica, one concentration, it is obvious that having singletons increases both the false positives and the false negatives ratio. Therefore, a minimum of close analogues would help in deciding if positive (or negative) results are true values simply by looking at the whole activity range across the cluster. At the same time – evidently – the higher the number of cluster members, the lower the "chemical diversity", no matter how this can be defined. To improve to some extent the cluster representativeness without limiting too much the diversity, the relationships between super and sub-scaffolds have been defined. In this way, it is possible to consider as belonging to the same cluster compounds that in accordance with other definitions should fall in different clusters. Furthermore, by applying Jarvis-Patrick[6] or Ward[7] algorithms using Tanimoto[8] distance and ECFP_6 descriptors[9] as implemented in Pipeline Pilot (Accelry), chemical diversity is measured over the whole structure, thus considering side chains as well as ring systems.

Once all compounds were selected and clustered, we faced the complexity of structuring a chemical library for efficient HTS. Our strategy to further facilitate the identification of preliminary SARs from HTS hits was to efficiently plate the IDDN collection. To do so, the IDDN library was divided into 5 sets, namely four "Diverse" and one "FollowUp" sets. "Diverse" sets were those collecting centroids and "pseudo-centroids" from each cluster, while the FollowUp set contained all the remaining selected compounds organized by clusters, as discussed in the forthcoming sections. While each of the four "Diverse" sets contained one representative for each cluster, being it the centroid or else the pseudo-centroids, in the "FollowUp" set compounds belonging to the same structural clusters were plated together, thus allowing expansion of hits identified during screening assays, without the need of "cherry picking". This last set, containing ~160K molecules, was designed to perform the preliminary SAR expansion of hits identified during HTS of the Diverse Sets. We thus designed a well-organized scheme that efficiently grouped structurally related chemotypes into the same plate. The algorithm is described in detail in the Methods, in brief for each vendor and for each cluster, it assigns each compound to a specific well in a specific plate until the plate is filled with compounds very similar by definition. This allows the use of each plate as a sort of cartridge containing highly enriched groups of compounds similar to those identified as hits in the first HTS step. Keeping in mind that the first HTS step is usually performed on the four Diverse Sets, as will be described in the Results section, this means that in the first phase four representatives of each cluster are tested, and in the second phase from six to about forty compounds of the same cluster are tested, with an average of fourteen. This is indeed a key innovative aspect of our scheme for the construction of a diverse chemical library for efficient HTS.

As a result, a well-organized and diverse chemical collection of about 200,000 compounds was built, with the major advantage of allowing a preliminary SAR of potential HTS hits. Remarkably, the final IDDN chemical library shows an optimal drug-like profile, highlighting the quality of the included compounds (see results section). Theretofore, the proposed scheme for the creation of an organized chemical library could in principle be used for larger collections, as well, helping in the creation of other chemical collections for efficient HTS campaigns.

6.2 METHODS

6.2.1 Initial compounds collection

We initially collected about 14 million compounds from many vendors. File cleaning and uniqueness checks were performed using standard SD file parsing software (Pipeline Pilot, Accelrys). Vendor files were submitted to removal of empty records and standardization of indexing fields, e.g. vendor name and catalogue number. Then, an internal identification number (GUID) was assigned to each compound. Other common file manipulations were performed, namely chemical groups standardization (e.g. Nitro, Sulfones etc), checks for consistency for protective groups (i.e. abbreviations corresponding to chemical structures), tautomer fixing, salt removal, and charge neutralization. Once a set of consistent chemical structures was obtained, the uniqueness check was performed. Then, the Canonical Smiles[10], the InChi and InChiKey were generated for each compound. Finally, compounds were filtered to remove undesired molecular structures. To do so, several filters were applied, using physico-chemical and topological properties, therefore removing unsuitable chemical moieties (PAINS[11] and Frequent Hitters[12,13,14]).

After this key step, the number of compounds was reduced to 3 million; suppliers were contacted to check whose compounds were resulting in 1.1 million available compounds that can be acquired.

This final set was considered for the additional steps for the IDDN library construction (Figure 49).

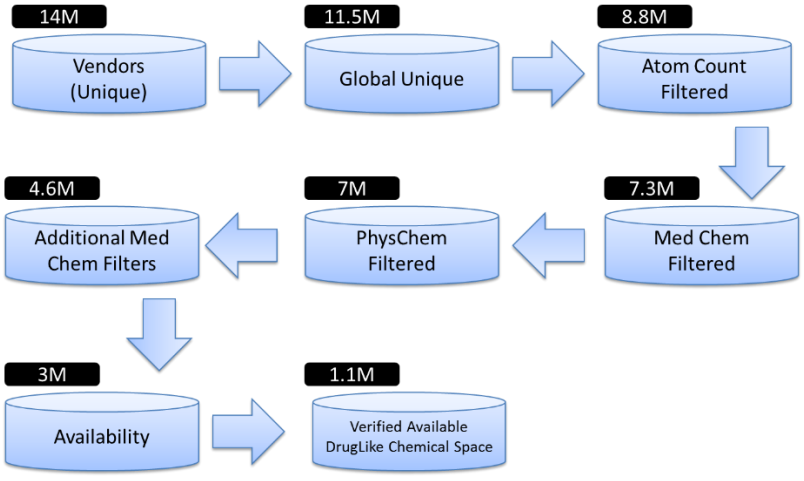


Figure 49. Drug-like filters cascade to identify the reference chemical space from the bulk collection of vendor's compounds

6.2.2 Compound clusterization

To build the final set of compounds (about 200K), we introduced a clusterization procedure to maximize the diversity among classes of compounds, using a novel scaffold-based metric called Molecular Assemblies (see chapter 5).

Taking into account one ring more or one ring less of the Decorated Scaffold, as a sub set of the Molecular Assemblies descriptor this reduced representation is still a hierarchical scaffold metrics. This useful representation allows us to consider the additional ring as a substituent, and the corresponding molecules as part of the same cluster (extended scaffold), because of the close similarity with Medicinal Chemists methods of classification.

Furthermore, since preliminary SAR from HTS data was one of the prerequisites, a compound description and classification in accordance with SAR in agreement with the Medicinal Chemistry "common sense" was mandatory, in order to obtain not only hits, but rather chemical classes ready to be moved toward the "Hit to Lead" phase.

Importantly, we employed a scheme to structure the IDDN chemical library in an efficient manner for a preliminary SAR expansion after HTS and data analysis. Toward this aim, scaffold decompositions were used to implement the concept of sub- and super- scaffold. Accordingly, a sub-scaffold is part of the parent scaffold lacking at least one ring system, while a super-scaffold has at least one ring system more than the parent one (Figure 50)

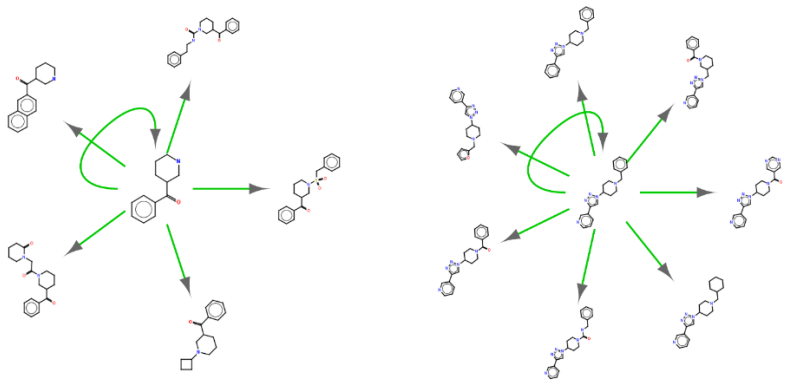


Figure 50.

Given this definition, the minimum number of rings to be considered is two. Therefore the procedure implemented starts with the extraction of Decorated Scaffolds as defined previously from all the selected compounds (~1M), then all the possible fragments containing at least two rings are recursively generated from each molecular framework. For example, if a given compound contains 4 rings, it will produce 3 fragments containing 2 rings each and 2 fragments containing 3 rings each. Of all these fragments, to avoid building unrealistic clusters, fragments non-existing within the collected structures are removed and only those really present in the scaffold pool.

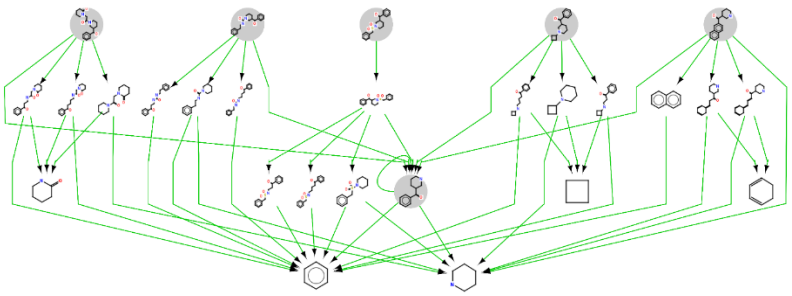


Figure 51. Cytoscape plot of the fragment decomposition of 5 Decorated Scaffolds according to Molecular Assemblies set of roles. Grey cycles indicates real scaffolds present in the dataset. All the upper 5 Decorated Scaffolds have one fragment in common that is itself a Decorated Scaffold. This pool of Decorated Scaffolds is defined as Extended Scaffold an all the molecules belonging to one of this Decorated Scaffolds are part of the same cluster (Figure 52).

It is important to underline that super-scaffold - i.e. molecular frameworks with one ring more than the reference compound (centroid) for any given cluster - are not built by adding ring systems, but, more simply, by establishing relationships within existing compounds. Molecules having a scaffold that is a sub- or a super- scaffold of another molecule scaffold have been grouped together and the new group is called extended scaffold (Figure 52).

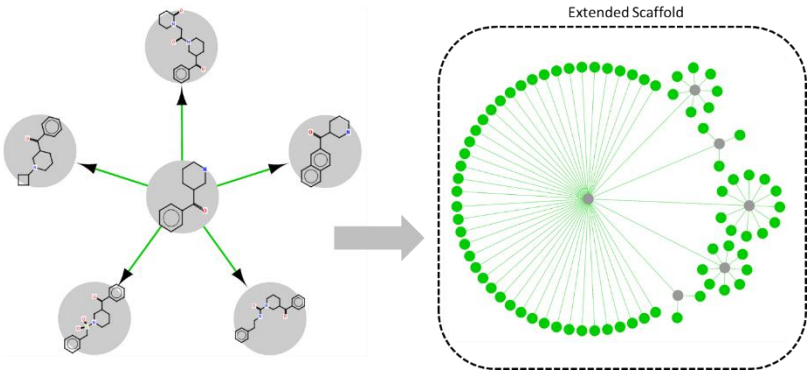


Figure 52. Molecules (Green) having a scaffold belonging to the same extended scaffold

Molecular Assemblies concept can also be very useful for establishing relations between different extended scaffolds (Figure 53).

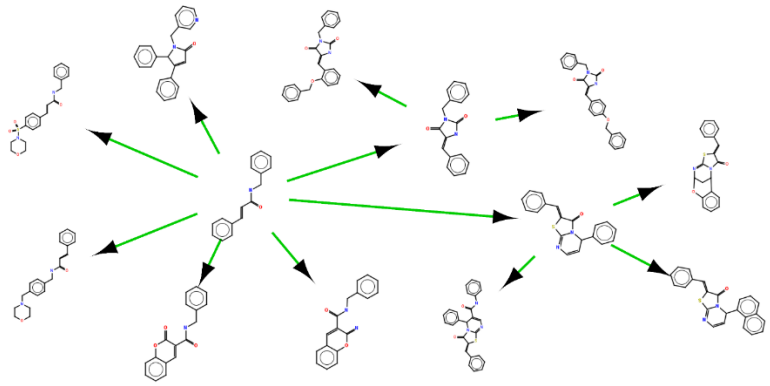


Figure 53.

Other Molecular Assemblies representations can be used to support the analysis of the chemical space and in the SAR analysis. For example, selecting one Basic Wireframe is possible to visualize all the compounds having that MABW in common and the structural relations connect all the molecules in the cluster (Figure 54).

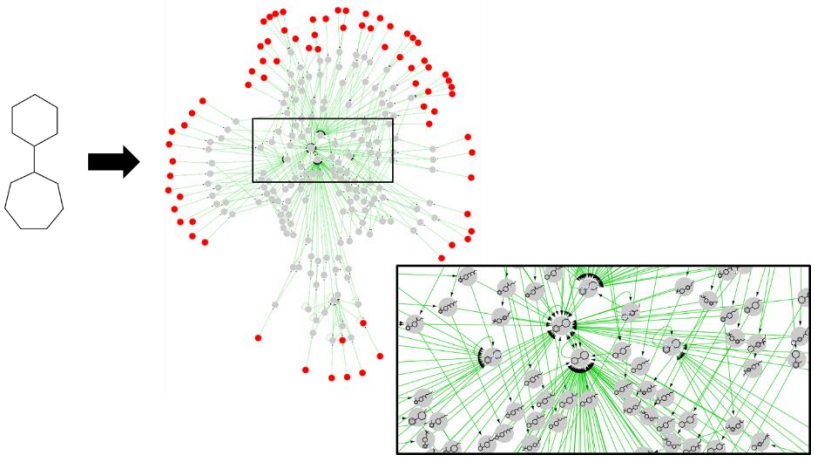


Figure 54.

Finally using InChi Keys string pattern, matching the associations between initial molecular framework and sub-scaffolds sharing part of the so-called Super Scaffold are established and the cluster is therefore defined.

The first analysis of the resulting clusters showed the presence of some overpopulated clusters (>1000 members). We decided to reduce the maximum number to 150, by selecting the corresponding most diverse molecules using Tanimoto ([8]) metric based of ECFP_6 descriptors.

To have an acceptable representation of different substitutions for molecules belonging to the same clusters in a collection designed to have 208K molecules we arbitrarily reputed that an average of 20 molecules for each cluster (i.e. 1/10000 of the whole set) would be sensible. At the same time the maximum number for each cluster was limited to 50 members. These two values (AvgSize 20, MaxSize 50) are the result of an iterative process so that the maximum number of clusters (i.e. scaffolds) available within the molecule dataset is selected, while the total number of compounds does not exceed 208K. At the same time the lower limit for each cluster was set to 10 to allow the selection of one representative for each Diversity Set still leaving six molecules for the hit expansion.

To choose 50 molecules out of – in some cases – 150 we again applied the diversity method based on Tanimoto ([8]) distance metric using ECFP_6 descriptors as implemented in PipelinePilot (Accelrys). The desired number of the most diverse molecules was selected and five non-redundant subsets were formed. The first four "Diverse" sets are made of representative members of each cluster obtained through the procedure described above. The first Diverse set collects all the cluster centroids, i.e. the molecules with the minimal distance from each other member within the same cluster. This was done using ECFP_6, Jarvis-Patrick algorithm and Tanimoto distance metric as implemented in Pipeline Pilot, resulting in 10000 compounds. The second, third and fourth Diverse sets consist of three other centroids, which are defined as follows: a) for each cluster as defined above, up to the top 50 most diverse molecules are selected; b) each cluster is then divided in three sub-clusters of about same size, using diversity calculated above as ranking parameter; c) out of each sub-cluster the centroid is identified; d) each centroid is then numbered and assigned to its own sub-sets. In the end four sets are produced: Cluster Centroids (LibraryID1), Sub-sets 1, 2 and 3 (LibraryID2, LibraryID3 and LibraryID4) of about 10000 compounds each for a total of 40000 compounds. The remaining ~168K molecules went into the fifth set. All and each of the molecules belonging to this fifth set must have one representative in each of the four diversity sets. Relationships between molecules are kept using InChikeys and those not fulfilling this criteria are removed.

6.3 RESULTS AND DISCUSSION

As briefly previously reported, the two main goals of this library construction were a) capability to derive a preliminary SAR from HTS data and b) assembly of the library so that the above goal is possible without performing cherry picking, i.e. selecting single compounds, but otherwise using complete plates as "scaffold collections". To achieve these, some conditions were fulfilled. The first one is that each scaffold, or chemotype, should have enough representatives with typical substituent variations, but still maximizing diversity[15,16]. Then the classification should take care of relationships between similar scaffolds, in particular those connected by addition or removal of a ring system. Finally, similar compounds must be plated together, so that the ratio of interesting compounds versus the other plate members is as high as possible, once a scaffold has been found active after screening of the four diversity sets. Since the compounds came from six different vendors, the approach took in account this constraint as well, and, for obvious logistical reasons, compounds coming from different vendors are not intermixed, and there are no duplications between vendors.

From the above requirements, we derived all the parameters we used in the library design. Given 208000 compounds as the maximum number to be taken into account, in accordance with logistic and economical constraints, we decided that an average of 20 compounds for each cluster could produce a minimal SAR. At the same time this produces what is considered the minimal diversity set (~10000 cmpds), which is at the same time the number of clusters based on the scaffold metrics described above. Considering that these constraints could dramatically reduce chemical diversity, and keeping in mind the definition of Decorated Scaffold, where the scaffold is almost identical within the clusters, we decided to select at every step where it was required always the top diverse molecules, since the diversity would have been contained in the side chains and therefore would not disrupt the scaffold based clusterization. Furthermore we decided to replicate for a total of four times the process of centroids extraction for each cluster, to provide another degree of flexibility when performing HTS assays. This resulted in 4 diversity subsets, each of them containing about 10000 molecules, each of them representative of the whole library, and, when used all together, representing 20% of the whole library.

This new definition was implemented because including molecules having a fragment scaffold of the super scaffold could simplify the identification of the part of the scaffold carrying the most relevant pharmacophore; in addition, introducing molecules having scaffolds that can create relations between different scaffolds having a partial overlap in the chemical structure allows a better understanding of the role of the scaffold and the effects of scaffold modifications in the biological activities helping in the identification of analogues and, eventually, for scaffold hopping activities.

6.4 CONCLUSIONS

In this work, we have demonstrate how is possible to implement Molecular Assemblies in the library design process in order to generate drug like, SAR-oriented, HTS ready chemical libraries. The definition of an extended scaffold, as an application of the fragmentation's rules, overcome the limits of other scaffold metrics that are not able to cluster together molecules having only a partial overlap in the scaffold because of the intrinsic definition of scaffold (the rings and the chains connecting them).

The use of the full MA descriptor, instead of extended scaffold, will guide the analysis of the results offering multiple interpretation of the SAR. Active molecules may have in common the scaffold, the framework of

the wireframe or even a fragment of them in any case MA descriptor will able to generate all possible relations between active molecules.

The use of MA metrics for clustering enabled the efficient clustering of the entire chemical space. MA have no need to introduce bias such as the expected number of molecules per cluster, the expected number of clusters, the number of nearest neighbor, the definition of cluster centroid.

As demonstrated in Chapter 5, cluster by MA is a simple function of the number of molecules, so it is possible to cluster the initial set of 1.1 million compounds in a few minutes. It is a remarkable result if compared to other clustering methods based on distances were clustering is a function of the number of molecules to the power or two or even three in the case of Jarvis-Patrick algorithm.

The design and acquisition of this chemical library is a shining example of a successful attempt to explore novel R&D models. The Italian Drug Discovery Network, composed of pharmaceutical companies, public and private research institutions, as a small precompetitive joint initiative, has demonstrated that is possible to accelerate and promote drug discovery. Providing, as a network, the human and financial resources to achieve the required critical mass to compete in this arena.

6.5 References

1. Mayr, L. M.; Bojanic, D., Novel trends in high-throughput screening. *Current opinion in pharmacology* **2009,** *9* (5), 580-8.
2. Blaxill, Z.; Holland-Crimmin, S.; Lifely, R., Stability through the ages: the GSK experience. *Journal of biomolecular screening* **2009,** *14* (5), 547-56.
3. Beresini, M. H.; Liu, Y.; Dawes, T. D.; Clark, K. R.; Orren, L.; Schmidt, S.; Turincio, R.; Jones, S. W.; Rodriguez, R. A.; Thana, P.; Hascall, D.; Gross, D. P.; Skelton, N. J., Small-Molecule Library Subset Screening as an Aid for Accelerating Lead Identification. *Journal of biomolecular screening* **2014,** *19* (5), 758-770.
4. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **2001,** *46* (1-3), 3-26.
5. Bruns, R. F.; Watson, I. A., Rules for identifying potentially reactive or promiscuous compounds. *Journal of medicinal chemistry* **2012,** *55* (22), 9763-72.
6. Jarvis, R. A.; Patrick, E. A., Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans. Comput.* **1973,** *22* (11), 1025-1034.
7. Ward, J. H., Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **1963,** *58* (301), 236--244.
8. Rogers, D. J.; Tanimoto, T. T., A Computer Program for Classifying Plants. *Science* **1960,** *132* (3434), 1115-8.
9. Rogers, D.; Hahn, M., Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010,** *50* (5), 742-54.
10. (a) Weininger, D., SMILES. 3. DEPICT. Graphical depiction of chemical structures. *Journal of Chemical Information and Computer Sciences* **1990,** *30* (3), 237-243; (b) Weininger, D., SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988,** *28* (1), 31-36; (c) Weininger, D.; Weininger, A.; Weininger, J. L., SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences* **1989,** *29* (2), 97-101.
11. Baell, J. B.; Holloway, G. A., New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry* **2010,** *53* (7), 2719-40.

12. (a) Schorpp, K.; Rothenaigner, I.; Salmina, E.; Reinshagen, J.; Low, T.; Brenke, J. K.; Gopalakrishnan, J.; Tetko, I. V.; Gul, S.; Hadian, K., Identification of Small-Molecule Frequent Hitters from AlphaScreen High-Throughput Screens. *Journal of biomolecular screening* **2013,** *19* (5), 715-726; (b) Aronov, A. M.; Murcko, M. A., Toward a pharmacophore for kinase frequent hitters. *Journal of medicinal chemistry* **2004,** *47* (23), 5616-9; (c) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E. M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjogren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G., Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *Journal of medicinal chemistry* **2002,** *45* (1), 137-42.
13. Brown, N.; Zehender, H.; Azzaoui, K.; Schuffenhauer, A.; Mayr, L. M.; Jacoby, E., A chemoinformatics analysis of hit lists obtained from high-throughput affinity-selection screening. *Journal of biomolecular screening* **2006,** *11* (2), 123-30.
14. Faller, B.; Wang, J.; Zimmerlin, A.; Bell, L.; Hamon, J.; Whitebread, S.; Azzaoui, K.; Bojanic, D.; Urban, L., High-throughput in vitro profiling assays: lessons learnt from experiences at Novartis. *Expert opinion on drug metabolism & toxicology* **2006,** *2* (6), 823-33.
15. Krier, M.; Bret, G.; Rognan, D., Assessing the scaffold diversity of screening libraries. *Journal of chemical information and modeling* **2006,** *46* (2), 512-24.
16. Langdon, S. R.; Brown, N.; Blagg, J., Scaffold diversity of exemplified medicinal chemistry space. *Journal of chemical information and modeling* **2011,** *51* (9), 2174-85.

Development and Validation Of In Silico Tools For Efficient Library Design and Data Analysis in High Throughput Screening Campaigns

Chapter 7: Concluding Remarks

7 CONCLUDING REMARKS

I have dedicated my PhD to the development and validation of new methodologies and software tools for research and development workflow. I have also been able to apply such tools to actual drug discovery projects to demonstrate that they were able to implement and support the very first step of R&D processes. In fact, hit finding and hit to lead development are the crucial steps for the fate of the projects, even though they are also the cheaper ones. The first identification of high quality hits, that can be rapidly progressed to validated hit series and then optimized to obtain good preclinical leads, strongly bias the chance of success of the projects. In fact, an ineffective choice of the starting point will produce a failure in the subsequent and more costly phases.

The two methodologies I identified are:

- *LiGen*, a molecular mechanics engine for structure-based virtual screening
- *Molecular Assembles*, a molecular descriptor for cheminformatics applications

Even though they are very different from each other, they both find application in the same computational field. They could be used in an integrated way thus addressing the whole challenge of identifying valuable hits, instead of supporting vertically specific tasks one by one with standalone computational tools.

In the specific case of this thesis, Molecular Assembles has been used in the design and assembly of the IDDN chemical library (chapter 6) that was subsequently tested through biologic functional assays. In these processes, *Molecular assembles* makes HTS data analysis easier by guiding the SAR analysis. Moreover, the use of well-designed libraries in structure based virtual screening allows the optimization of computational resources and human effort in the analysis of the results, since they are immediately focused on high quality drug-like molecules with a good balance between chemical diversity and the possibility to establish structure activities relationships.

On the other hand LiGen has proved able in supporting the de novo design of novel C5aR allosteric inhibitors.

Such a combined approach greatly enhances the expected hit rate. Nevertheless, even more important than the hit rate is the possibility to analyze the results of small congeneric series (molecules having the same scaffold) by the means of the 3D interactions with the active site. This *modus operandi* allows the validation of the prediction power of the 3D model with the experimentally determined ligand based SARs directly in the first round of screening.

The final goal of any HTS campaign is the identification of a small set (from 3 to 10) of validated hit series having the highest chances to progress up to a lead instead of tens of single active compounds. So the final goal of computational medical chemistry and more in general of the CADDD is to foster the integration of the in silico tools with the drug discovery and development workflow providing reliable predictions by using interpretable models to support the decision making process.

Molecular Assembles and LiGen are ongoing projects, used in all our early drug discovery projects. Changes and optimizations are determined by the performance and the limitations in real world applications. For example, working on a project with very flexible ligands, we revealed the need for an implemented novel algorithm for the identification of the first pose of the ligand in the active site.

The main enhancements already scheduled are:

- Molecular Assembles will be re-written in C++ (the actual implementation is in Java and requires Accelrys Pipeline Pilot API) using the part of the LiGen code devoted to the substructure search and reagent mapping. This implementation will also allow LiGen natively to take into account the scaffold metrics in the virtual screening.
- LiGen scoring function will be re-implemented form scratch and in 2015 the optimization of the source code for the INTEL xeon phi accelerators will be completed.

8 TABLE OF CONTENTS