

<b>Institution:</b> University of Sheffield
<b>Unit of Assessment:</b> 36 - Communication, Cultural and Media Studies, Library and Information Management
<b>Title of case study:</b> More cost-effective drug discovery using virtual screening
<b>1. Summary of the impact</b> <p>The discovery of a new drug can take 10 years, cost in excess of one-billion dollars and involve synthesising and testing thousands of possible drug molecules. Virtual (i.e., computer-based) screening is used in the early stages of drug discovery to focus attention on those molecules in a chemical database that are most likely to exhibit the required drug action and that are hence priority candidates for further, more detailed study. Virtual screening thus increases the cost-effectiveness of pharmaceutical research by bringing novel drugs to patients more quickly.</p> <p>Work in Sheffield since 1993 on virtual screening has resulted in three computer programs that enable much more effective screening to take place than was previously possible and that are now used throughout the world-wide pharmaceutical industry: GALAHAD (Genetic Algorithm with Linear Assignment for the Hypermolecular Alignment of Datasets), GASP (Genetic Algorithm Superimposition Program) and GOLD (Genetic Optimization for Ligand Docking).</p>
<b>2. Underpinning research</b> <p><b>Background</b></p> <p>Two of the most important virtual screening techniques are <i>ligand docking</i> and <i>pharmacophore mapping</i>. These both focus on the three-dimensional (3D) structures of molecules and are computationally demanding, especially when molecules are <i>conformationally flexible</i>, i.e., can adopt a range of 3D shapes when they interact with a biological macromolecular target, typically a specific protein in the body.</p> <p>Ligand docking (the focus of GOLD) involves identifying those molecules in a database that fit a target, in much the same way as a key fits into a lock, since those with a good fit may be biologically active. Pharmacophore mapping (the focus of GASP and GALAHAD) involves identifying the structural features common to molecules that have already been shown to be biologically active.</p> <p><b>Sheffield research</b></p> <p>Docking and pharmacophore mapping with flexible molecules are examples of combinatorial optimization problems for which efficient, conventional algorithms are not available. However, they are very well suited to genetic algorithm (GA) techniques, the use of which in chemoinformatics was pioneered in Sheffield work in the early Nineties and which form the basis of GOLD, GASP and GALAHAD.</p> <p>The GA in the Sheffield GOLD program identifies the best possible fit of a flexible molecule into a target protein. This was an entirely novel approach when the prototype program was first described in 1995 [R1], following a collaboration with Wellcome (now GlaxoSmithKline, or GSK) by Willett (Professor of Information Science since 1991). Extensive development and testing [R2] in an MRC/DTI LINK project at Sheffield (£205K, 1995-97) resulted in a program that was rapidly adopted by industry once it was distributed by the Cambridge Crystallographic Data Centre (CCDC, the industrial partner in the LINK project).</p>

The GA in the Sheffield GASP program is analogous to that in GOLD: the latter tests the fit of a molecule to a protein while the former tests the fit of a series of active molecules to each other [R3]. GASP was developed during the GOLD project and was subsequently commercialized by a US chemoinformatics company, Tripos Inc. (St Louis, MO) to complement an existing product, which was based on Sheffield research in the late Eighties and which was less well-suited than GASP to the handling of flexible molecules. A subsequent funded collaboration between Willett and Tripos (£121K, 2002-04) led to a novel method for aligning the 3D structures of pairs of molecules [R4] that lies at the heart of a further pharmacophore program, GALAHAD [R5], that Tripos now distributes. GALAHAD also makes use of a multiple objective GA, an approach to optimization that was first applied to chemoinformatics in a collaboration with GSK (£129K, 1999-2002) by Gillet (then Senior Lecturer and Professor of Chemoinformatics since 2009) [R6]. This approach has since been successfully demonstrated in a range of other chemoinformatics applications.

The high quality of the research carried out by the Sheffield chemoinformatics group (Gillet, Willett and Holliday (Senior Research Manager since 1999)) is further evidenced by:

- Inspection of the core journal for chemoinformatics, the *Journal of Chemical Information Modeling*, where many articles are published by industrial, rather than academic, researchers and where the Sheffield group has provided more contributions than any other organization over the journal's 53-year lifetime.
- The Sheffield group being the recipients of 2012 Jason Farradane Award of the UK Electronic Information Group "in recognition of outstanding work in the information profession", the commendation highlighting the "joint research projects with many of the world's leading pharmaceutical, agrochemical and software companies".
- Willett being the only person to have been the recipient of all three of the American Chemical Society's awards for contributions to chemoinformatics and the pharmaceutical industry: the 1993 Herman Skolnik Award, the 2005 Award for Computers in Chemical and Pharmaceutical Research, and the 2010 Patterson-Crane Award.

The group's publications since 1993 have attracted over ten-thousand citations in the *Web of Science* (WoS), with many of these reflecting the impact of the work on industrial research. Multi-national pharmaceutical companies that have cited Sheffield research at least 20 times include Abbott, AstraZeneca, Bristol-Myers Squibb, Eli Lilly, GSK, Merck, Novartis, Pfizer and Roche. The work has also been extensively cited by the world's major chemoinformatics software and database companies, with such companies citing Sheffield work at least 5 times including Accelrys, BCI, CCDC, CCG, CEREP, De Novo, Leadscape, Molecular Networks, OpenEye, Schroedinger and Tripos. As [S1] notes "*The group can boast a prodigious corpus of published research as well as a number of prestigious alumni in industry and academia, with the result that most serious researchers in the field have had some direct or indirect contact with them...They also provide much-needed leadership to chemoinformaticians - in the pharmaceutical industry as well as in the academic community - by maintaining exceptional standards of objectivity and critical examination of their own work that are typically weaknesses in our field of research*".

### 3. References to the research

- R1. Jones, G., Willett, P. & Glen, R.C. "Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation". *Journal of Molecular Biology*, **245**, 1995, 43-53. (5-year journal impact factor IF =3.89, 691 citations in *Web of Science*)
- R2. Jones, G., Willett, P., Glen, R.C., Leach, A.R. & Taylor, R. "Development and validation of a genetic algorithm for flexible docking." *Journal of Molecular Biology*, **267**, 1997, 727-748. (IF=3.89, 2350 citations)

- R3. Jones, G., Willett, P. & Glen, R.C. "A genetic algorithm for flexible molecular overlay and pharmacophore elucidation". *Journal of Computer-Aided Molecular Design*, **9**, 1995, 532-549. (IF=3.61, 246 citations)
- R4. Richmond, N.J., Willett, P. & Clark, R.D. "Alignment of three-dimensional molecules using an image recognition algorithm". *Journal of Molecular Graphics and Modelling*, **23**, 2004, 199-209. (IF 2.22, 36 citations)
- R5. Richmond, N.J., Abrams, C., Wolohan, P.R.N., Abrahamian, E., Willett, P. & Clark, R.D. "GALAHAD. 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D." *Journal of Computer-Aided Molecular Design*, **20**, 2006, 567-587. (IF=3.61, 80 citations)
- R6. Gillet, V.J., Khatib, W., Willett, P., Fleming, P.J. & Green, D.V.S. "Combinatorial library design using a multiobjective genetic algorithm." *Journal of Chemical Information and Computer Sciences*, **42**, 2002, 375-385. (IF=4.07, 99 citations)

#### 4. Details of the impact

The impact of the Sheffield chemoinformatics work in terms of increasing economic prosperity for the pharmaceutical industry and increasing the range of medicines available to patients has been two-fold: by means of software developed in Sheffield and then distributed to industry; and by means of extensive industrial adoption of algorithms and methods published by the group in the open literature. The reach of the work is demonstrated by its current use throughout the pharmaceutical industry and the significance by its role in helping to discover new drugs that can benefit patients throughout the world.

GOLD has been distributed by CCDC since 1998 and is currently used for pharmaceutical research in over 50 different countries:

- "For many years, GOLD has been one of the top two protein-ligand docking programs for industrial and academic research in drug discovery, and is used by around 60 major pharmaceutical companies and more than 600 Universities worldwide" [S2];
- "My group of 18 modellers and chemoinformaticians are actively involved in the support of many drug discovery projects; our work has led to the discovery of several novel chemical entities now in the clinic or on the market. One of the tasks we perform is virtual screening of large databases of chemical structure, and for that purpose we regularly use GOLD as the tool of choice" [S3].

GASP has been distributed by Tripos since 1996, with an enhanced version available since 2003, and GALAHAD has been distributed by them since 2005. These three products have been very successful commercially, with total sales to date of ca. £7 million and with royalties to the University of Sheffield of ca. £900K (£82K in 2008-12). The software licensees include many of the world's leading chemical companies, with, e.g., GASP and GALAHAD users including Amgen, Bayer, Boehringer Ingelheim, Bristol-Myers Squibb, DuPont, Eli Lilly, Genentech, GSK, Novo Nordisk, Sanofi Aventis and Syngenta *inter alia*.

License holders such as these have made very extensive use of the software, as demonstrated by published accounts in the scientific literature of the use of GALAHAD, GASP and GOLD in companies' internal drug-discovery projects. WoS searches in October 2013 for the six references listed in Section 3 reveals over 250 citations to this research by industrial companies in 2008-13. For example, articles in two leading medicinal chemistry journals (*Journal of Medicinal Chemistry* (IF=5.38) and *Bioorganic and Medicinal Chemistry* (IF=3.15)) describe the use of the Sheffield software by world-leading pharmaceutical companies to support research in cancer (Genentech), heart disease (Pfizer and Proctor & Gamble), cellular signalling (GSK), cognitive impairment such as schizophrenia and Alzheimer's (Abbott), and obesity (Takeda) *inter alia*.

The impact of the Sheffield work is much broader than just the three specific commercial products, in that techniques described by the Sheffield group in the open literature have been widely adopted, as evidenced both by literature citation and by incorporation in software that is used around the world on a daily basis. For example:

- “We also follow the work of the Sheffield group keenly and we have adopted some of their ideas on data fusion techniques and multi-objective optimization into our own code; we have done this more with the research coming out of this department than we have with any other academic cheminformatics group in Europe “ [S3];
- “Sheffield research has been at the heart of all the commonly used, indispensable systems for handling chemical reactions, and 2D, 3D and generic chemical structures in industry.....Very few universities worldwide are capable of addressing the “virtual R&D” problem; Sheffield is the leader” [S4].

#### 5. Sources to corroborate the impact

- S1. Email from Director of Life Sciences, Simulations Plus corroborates impact of Sheffield cheminformatics work on the pharmaceutical industry in general.
- S2. Email from Emeritus Research Fellow, Cambridge Crystallographic Data Centre corroborates reach of Sheffield cheminformatics work in the pharmaceutical industry.
- S3. Email from Global Head of Computer-Aided Drug Discovery, Novartis corroborates use of Sheffield cheminformatics work and corporate use of GOLD at Novartis.
- S4. Email from Associate Editor, *Journal of Chemical Information and Modeling* corroborates impact of Sheffield cheminformatics work on the pharmaceutical industry in general, and has detailed knowledge of the contribution to the *Journal of Chemical Information and Modeling*.
- S5. Director of Computational Chemistry Europe, GlaxoSmithKline. Can corroborate impact of Sheffield cheminformatics work on the pharmaceutical industry in general, and has detailed knowledge of the development of GOLD.