Institution: University of Sheffield		
Unit of Assessment: D-34 Communication, Cultural and Media Studies, Library and Information Management		
Title of case study: Enhancing the effectiveness of drug discovery programmes		
Period when the underpinning research was undertaken: 2003–2015		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
V.J. Gillet	Professor	2003–present
P. Willett	Professor	2003–2019
Period when the claimed impact occurred: 2014–2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact (indicative maximum 100 words)

The case study describes two computer-based applications that use the concept of molecular similarity to enhance the effectiveness of drug discovery programmes, with the research having impact in the area of health and well-being. One application helps to provide drugs for people suffering from rare diseases, and the other helps pharmaceutical companies discover novel drugs. The research has enhanced the effectiveness of pharmaceutical research, where the discovery of a new drug can take 10 years, cost in excess of $1 billion and involve synthesising thousands of possible drug molecules.

2. Underpinning research (indicative maximum 500 words)

The study describes two applications of the concept of molecular similarity, i.e. how structurally alike the computer representations of chemical molecules are to each other, to assist in bringing novel drugs to the general public. One application, developed in 2010-15 in a collaboration with the European Medicines Agency (EMA), applied molecular similarity to the licensing of drugs to treat very rare diseases. The other, developed in 2003-09 in collaboration with two of the world's major drug companies, GlaxoSmithKline (GSK) and Sanofi, used similarity calculations based on a novel type of computer representation, "reduced graphs", to assist in the discovery of potential drug molecules.

Orphan drugs

Most new drugs are designed to treat diseases where there are sufficient sufferers to amortise the massive R&D costs of drug discovery, but rare diseases may affect insufficient patients to make conventional procedures economically viable. In such cases, a drug can be given "orphan drug" status if it can be shown that the new drug is markedly dissimilar to existing drugs for that rare disease. The EMA is the regulatory authority responsible for licensing drugs for use throughout the EU, and its Committee for Medicinal Products for Human Use (CHMP) evaluates 12-15 orphan-drug applications each year, with a successful application meaning that no similar compound can be registered and marketed to patients for a 10-year period.

The Sheffield research developed software to quantify the concept of "markedly dissimilar" [R1, R2], using a novel application of the "fingerprint", a simple computer representation that describes the atoms and bonds comprising the structure of a molecule and that has been extensively studied at Sheffield [R3]. The dissimilarity is computed between the fingerprints describing a new drug that is being proposed by a pharmaceutical company and those describing each of any existing drugs for that disease. This computation then uses a statistical model to decide whether the new and existing drugs are indeed markedly dissimilar, and hence whether the new drug is eligible for consideration for orphan drug status.

Reduced graphs

Fingerprints effectively identify structurally similar molecules in searches of large databases [R3] but are less effective at finding molecules that have similar bioactivities but that have different structures. Such molecules are very valuable in drug discovery since they allow novel, and hence patentable, regions of chemical space to be explored, and an alternative representation, the reduced graph (RG) has been studied for this purpose.

An RG represents compounds by nodes (groups of atoms and bonds) that are connected by the topological (through-bond) distances between atoms in the nodes (rather than by the individual atoms and bonds as in a fingerprint). The RG was originally developed in Sheffield to describe the structures in chemical patents, but was adapted for database searches to identify compounds with the same bioactivity but with different structures [R4]. A collaboration with GlaxoWellcome (now GSK) showed the RG approach to be effective at finding compounds that share the same activity but which would not be found using traditional fingerprints [R4]. Further collaborations with GSK [R5] demonstrated the effectiveness of RGs in identifying structure-activity relationships using machine learning techniques, and with Sanofi [R6] explored the use of RGs to suggest functional group replacements that can improve the properties of a potential drug in order to increase its potency or to reduce undesirable side-effects.

3. References to the research (indicative maximum of six references)

R1. Franco, P., Porta, N., Holliday, J. D., & Willett, P. (2014). The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation. *Journal of Cheminformatics*, *6*(1), 5. https://doi.org/10.1186/1758-2946-6-5

R2. Franco, P., Porta, N., Holliday, J. D., & Willett, P. (2017). Molecular similarity considerations in the licensing of orphan drugs. *Drug Discovery Today*, *22*(2), 377–381. https://doi.org/10.1016/j.drudis.2016.11.024

R3. Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, *11*(23–24), 1046–1053. https://doi.org/10.1016/j.drudis.2006.10.005

R4. Gillet, V. J., Willett, P., & Bradshaw, J. (2003). Similarity Searching Using Reduced Graphs. *Journal of Chemical Information and Computer Sciences*, *43*(2), 338–345. https://doi.org/10.1021/ci025592e

R5. Birchall, K., Gillet, V. J., Harper, G., & Pickett, S. D. (2008). Evolving Interpretable Structure−Activity Relationships. 1. Reduced Graph Queries. *Journal of Chemical Information and Modeling*, *48*(8), 1543–1557. https://doi.org/10.1021/ci8000502

R6. Birchall, K., Gillet, V. J., Willett, P., Ducrot, P., & Luttmann, C. (2009). Use of Reduced Graphs To Encode Bioisosterism for Similarity-Based Virtual Screening. *Journal of Chemical Information and Modeling*, *49*(6), 1330–1346. https://doi.org/10.1021/ci900078h

4. Details of the impact (indicative maximum 750 words)

Orphan drug research

The Sheffield research has impact for three types of beneficiary: international pharmaceutical companies developing novel drugs; the regulatory authority responsible for the evaluation and supervision of those drugs throughout the EU; and, most importantly, patients suffering from rare diseases who can be treated with drugs that would not otherwise be commercially viable.

The fingerprint software described above has been available for use by CHMP since 2014. A marked dissimilarity in structure is one of several criteria that are used by CHMP to decide whether a new drug differs sufficiently from any existing remedies and hence whether orphan-drug status is appropriate. The software provides CHMP with a quantitative basis for helping to make this vital decision; and in-house implementations of the published method [R1, R2] are also used by pharmaceutical companies when submitting new drugs for consideration by the EMA, e.g. Novartis (the world's sixth largest pharmaceutical company by revenue in 2019) [S1].

Three submissions to CHMP illustrate the use of the fingerprint approach [S2-S4]. Novartis Europharm Ltd. used nine different types of fingerprint to demonstrate that their potential orphan drug Rydapt was markedly dissimilar to three existing orphan drugs for the treatment of specific forms of myeloid leukaemia and mastocytosis. Tesaro UK used one particular type of fingerprint to demonstrate that their potential orphan drug Zejula was markedly dissimilar to two existing orphan drugs for the treatment of specific forms of ovarian, fallopian tube and peritoneal cancer; and this was confirmed by CHMP using five other types of fingerprint. CHMP used a range of fingerprints to demonstrate that a potential orphan drug Chenodeoxycholic acid sigma-tau (later renamed Chenodeoxycholic acid Leadiant) proposed by Sigma-tau Arzneimittel GmbH was not markedly dissimilar to an existing orphan drug for the treatment of inborn errors of primary bile acid synthesis due to sterol 27 hydroxylase deficiency, a finding that necessitated further investigation of the submission by the EMA. The three applications were subsequently awarded orphan drug status for use within the EU and most recently have been approved for use within the UK by the NHS to treat acute myeloid leukaemia, recurrent ovarian cancer and inborn errors of primary bile acid synthesis [S5].

Reduced graph research

The Sheffield RG research has impact on the software used to support early-stage R&D in industrial drug discovery, and the principal beneficiaries are hence international pharmaceutical companies and the patients needing drug treatments. This impact takes place via inclusion in corporate information systems and in products marketed by chemoinformatics software companies.

The software developed in collaboration with GSK [R4, R5] is now used by them for their routine compound acquisition procedures. These identify chemical compounds synthesised by external suppliers that should be purchased to increase the structural diversity of their corporate database of compounds [S6]. GSK also employs RGs for virtual screening, i.e., computational procedures for prioritising compounds for testing for biological activity, and as part of an Artificial Intelligence (AI) based system for identifying potential drug molecules. RGs also formed the

basis for the Extended Reduced Graph (or ErG). This was originally developed at Eli Lilly and ErGs are now available for use throughout industry as part of the RDKit module https://www.rdkit.org/) in GitHub, which is the world's largest repository of open-source software [S1].

Commercially products that have adopted the RG technology are available from Discngine and Lhasa Limited. The software developed in collaboration with Sanofi [R6] has been implemented at the life sciences software company, Discngine, under the name of "pharmacophore graph" (PG). Since August 2013, software components using PGs have been licensed to 14 organisations comprising biotechs, non-profits and major pharmaceutical companies (including five of the ten companies with the largest 2019 revenues) [S7]. Examples of research that use RG or PG software and published by pharmaceutical companies include the treatment of tobacco addiction (GSK), epigenetics studies (BioFocus) and the design of novel trypsin inhibitors (Abbott) [S8]. Lhasa Limited is a not-for-profit that distributes data and software for predicting molecular properties such as toxicology and metabolism to over 300 organisations worldwide to streamline the drug development process and to reduce animal testing. Among their products is Sarah Nexus, which uses an RG-based fragmentation process as part of a statistical software tool to identify potentially toxic chemicals and hence to reject unsuitable drug candidates. Sarah Nexus is used by more than 100 organisations globally including many large pharmaceutical companies, regulatory organisations and universities. In 2017 Sarah Nexus was recognised by the Queen's Awards in the Innovation category [S9].

5. Sources to corroborate the impact (indicative maximum of 10 references)

S1. Testimonial from Associate Director Data Science, Novartis on the use of similarity assessment in applying to the EMEA for orphan drug status at Novartis. Email trace: Novartis-Stiefl.pdf

S2. *CHMP Assessment Report for Rydapt, on similarity with histamine dihydrochloride (Ceplene), azacitidine (Vidaza), and decitabine (Dacogen)* - EMA report no. EMA/CHMP/729473/2016. In Section 2.3 of this report, the sub-section Applicant's Position describes the use by Novartis Europharm Ltd of a range of fingerprints to investigate the degree of similarity between Rydapt and three existing orphan drugs Ceplene, Vidaza and Dacogen.

S3. *CHMP Assessment Report for Zejula on similarity with Lynparza and Yondelis* - EMA report no. EMA/CHMP/851422/2016. In Section 2.3 of this report, the sub-section Applicant's Position describes the use by Tesaro UK Ltd. of one type of fingerprint to investigate the degree of similarity between Zejula and two existing orphan drugs Lynparza and Yondelis, and the sub-section CHMP Position describes the use by EMA of several additional fingerprints for this purpose.

S4. *CHMP Assessment Report for Chenodeoxycholic acid sigma-tau on similarity with Cholic acid, authorised orphan product* - EMA report no. EMA/CHMP/293064/2016. In Section 2.2 of this report, the sub-section CHMP Position describes the use by EMA of a range of fingerprints to investigate the degree of similarity between Chenodeoxycholic acid sigma-tau and the existing orphan drug Cholic acid.

S5. European public assessment reports for Rydapt, Zejula and Chenodeoxycholic acid Leadiant, including confirmation of their orphan-drug status, are at https://www.ema.europa.eu/en/medicines/human/EPAR/rydapt, https://www.ema.europa.eu/en/medicines/human/EPAR/zejula, and

https://www.ema.europa.eu/en/medicines/human/EPAR/chenodeoxycholic-acid-leadiant-previously-known-chenodeoxycholic-acid-sigma-tau; and confirmations of their availability in the UK for use by the NHS are at https://www.nice.org.uk/guidance/ta523/informationforpublic, https://www.globenewswire.com/news-release/2018/06/01/1515322/0/en/TESARO-Announces-Addition-of-ZEJULA-to-Cancer-Drugs-Fund-in-UK.html, and https://www.england.nhs.uk/commissioning/publication/cholic-acid-and-chenodeoxycholic-acid-for-treating-inborn-errors-of-bile-acid-synthesis-all-ages/

S6. GLAXOSMITHKLINE. A testimonial on the use of the reduced graph at GlaxoSmithKline by Scientific Director & GSK Fellow. GlaxoSmithKline-Pickett.doc.

S7. DISCNGINE. The Discngine Chemistry Collection is a commercially available software product in which one component called the Pharmacophore Graph is a “highly customisable reduced graph representation to perform lead hopping” - https://www.discngine.com/discngine-chemistry-collection. A testimonial on commercial uptake of their software has been provided by Vincent Le Guilloux.

S8. (a) Capelli, A. M., et al. (2011). Stable expression and functional characterization of a human nicotinic acetylcholine receptor with α6β2 properties: discovery of selective antagonists. *British Journal of Pharmacology, 163*(2), 313–329. https://doi.org/10.1111/j.1476-5381.2011.01213.x, (b) Ahrens, T. et al.. (2011). Efficient Hit-Finding Approaches for Histone Methyltransferases. *Journal of Biomolecular Screening, 17*(1), 85–98. https://doi.org/10.1177/1087057111422823, (c) Venhorst, J., Núñez, S., & Kruse, C. G. (2010). Design of a High Fragment Efficiency Library by Molecular Graph Theory. ACS Medicinal Chemistry Letters, 1(9), 499–503. https://doi.org/10.1021/ml100163s

S9. LHASA. Sarah Nexus is a software product developed by Lhasa Limited which includes an RG fragmentation method. A testimonial on the commercial uptake of Sarah Nexus has been provided by Head of Molecular Informatics and AI.