Institution: University of York		
Unit of Assessment: 8 - Chemistry		
Title of case study: Software for crystallography supporting new drug development		
Period when the underpinning research was undertaken: 2006-2020		
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:**
Kevin Cowtan	Professor	1995 to Present
Jon Agirre	Royal Society University Research Fellow	2013 to Present
Keith Wilson	Professor	1995 to Present
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)

Software for macromolecular crystallography developed at the University of York is widely used in both academia and industry, and is licensed to over 150 commercial sites worldwide. The software is used extensively by both global companies and SMEs in the pharmaceutical and biotechnology sectors. Economic impacts arise from software licensing, but much more significantly from the new pharmaceutical products the software helps create. Literature surveys show new drugs arising from fragment-based drug discovery pipelines since 2010 have overwhelmingly used York software in their discovery phases, leading to at least 17 identified clinical trials.

2. Underpinning research (indicative maximum 500 words)

York Structural Biology Laboratory (YSBL) is a unit of the Department of Chemistry at the University of York, founded in the 1970s by Professors Guy and Eleanor Dodson. The laboratory took a pioneering role in developing portable, general-purpose software for protein structure solution, where previously crystallographic studies often required custom software development. In the 1990s, computational scientists Professor **Wilson**, Dr Murshudov and Professor **Cowtan** joined the group and went on to develop new mathematical and computational methods, which enabled the solution of larger and more complex structures. Software produced at YSBL is distributed as part of the CCP4 suite [3.1] which includes a large number of York-developed components. York software has also been incorporated in other projects worldwide, such as the ISOLDE plugin for the UCSF ChimeraX software.

The success of the YSBL approach stems from the fact that it embodies a tradition of linking computational method developers with working crystallographers, the results of which are practical and useful tools for the whole community. In this context, YSBL has originated many world-leading crystallographic software applications, most notably COOT (Crystallographic Object-Oriented Toolkit), a graphical model building and validation software. COOT was initially developed in York by Emsley and **Cowtan** [3.2 cited >25,000 times], with further capabilities being continually added up to the present day [e.g. 3.3 cited >16,000 times]. A significant factor in the success of COOT has been the work that has gone into making its best-of-class model-building, refinement and validation tools available to the non-specialist user, as well as a powerful, interactive fitting tool. Recent developments include a growing library of tools for locating and modelling ligands – of particular relevance to pharmaceutical users, as well as support for electron cryo-microscopy (cryo-EM) data.

Cowtan pioneered important techniques for phase improvement and developed fast Fourier feature recognition, a method for identifying structural motifs in electron density maps. These techniques are heavily used in the BUCCANEER software for protein model building. A companion software package (NAUTILUS) has also been developed for the automated determination of nucleic acid structures [3.4]. **Cowtan** is the author of a powerful software library for crystallographic computation (CLIPPER), which has been extended by **Agirre** and others for

use with the PYTHON computer language, opening the door to a new generation of computational scientists.

More recently, **Cowtan**'s group has extended the BUCCANEER and NAUTILUS software packages for application to cryo-EM data. These software modules have been released to the community through the CCP-EM project (Collaborative Computational Project for electron cryo-microscopy), a UK partnership project supported by MRC. The group is now working on refinement of EM models using a new 'shift field' approach. **Wilson** led the development of two generations of graphical user interfaces for CCP4, improving accessibility, automation and data organisation. A focus on community accessibility has been a key that has driven the widespread use of the software in both academic and commercial organisations.

Cowtan and **Agirre** have also addressed the quality problems in deposited carbohydrate structures, implementing the PRIVATEER software [3.5] which has been used in the determination of most major medically relevant glycoproteins from 2016 to date, including SARS-CoV-2, ebolavirus and HIV-1. New restraint libraries and validation tools have been produced to reduce the error rate in future carbohydrate structures [3.6].

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

3.1. Winn, M.D., Ballard, C.C., **Cowtan, K. D**., Dodson, E.J., Emsley, P., Evans, P.R., ... & McNicholas, S.J. (2011) Overview of the CCP4 suite and current developments. Acta Crystallographica Section D: Biological Crystallography, **67**(4), 235-242. DOI: 10.1107/S0907444910045749
3.2. Emsley, P. & **Cowtan, K**. (2004) Coot: model-building tools for molecular graphics. Acta Crystallographica Section D: Biological Crystallography, 60, 2126-2132. DOI: 10.1107/S0907444904019158
3.3. Emsley, P., Lohkamp, B., Scott, W.G., & **Cowtan, K**. (2010) Features and development of Coot. Acta Crystallographica Section D: Biological Crystallography, **66**(4), 486-501. DOI: 10.1107/S0907444910007493
3.4. Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallographica Section D: Biological Crystallography, **62**(9), 1002-1011. DOI: 10.1107/S0907444906022116
3.5. Agirre, J., Iglesias-Fernández, J., Rovira, C., Davies, G.J., **Wilson, K.S**., & **Cowtan, K.D**. (2015) Privateer: software for the conformational validation of carbohydrate structures. Nature Structural and Molecular Biology, **22**(11), 833-834. DOI: 10.1038/nsmb.3115
3.6. Agirre, J., Davies, G. J., **Wilson, K.S.**, & **Cowtan, K.D**. (2015) Carbohydrate anomalies in the PDB. Nature Chemical Biology, **11**(5), 303. doi.org/10.1038/nchembio.1798

All references have been peer reviewed. [3.5] is being returned to REF2021.

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

COOT is widely used for crystallography in both academic and industrial contexts. It is typically cited in around *ten new publications every day*, and was cited in 95% of research papers reporting a major graphical model-building tool [5.4]. It plays a key role in state-of-the-art techniques in the pharmaceutical industry, by enabling scientists to rapidly interpret their data to understand how drug candidates interact with the molecules of life.

For every year from 2009 to 2016, the pharmaceutical industry spent more than any other section of UK industry on research, exceeding GBP4 billion per year and comprising ~20% of all R&D expenditure in UK business. In 2018, the BioIndustry Association reported that 'The UK has the strongest pipeline in Europe across all preclinical and clinical stages, with 351 preclinical products, 43 phase I, 70 phase II and 15 phase III' and 'UK biotech company IPOs raised more than twice as much money in 2017 (GBP234,000,000) than in 2016 (GBP105,000,000).' [5.1] YSBL software generates economic impact from its almost exclusive use in supporting these multi-billion GBP UK and international investments in drug development.

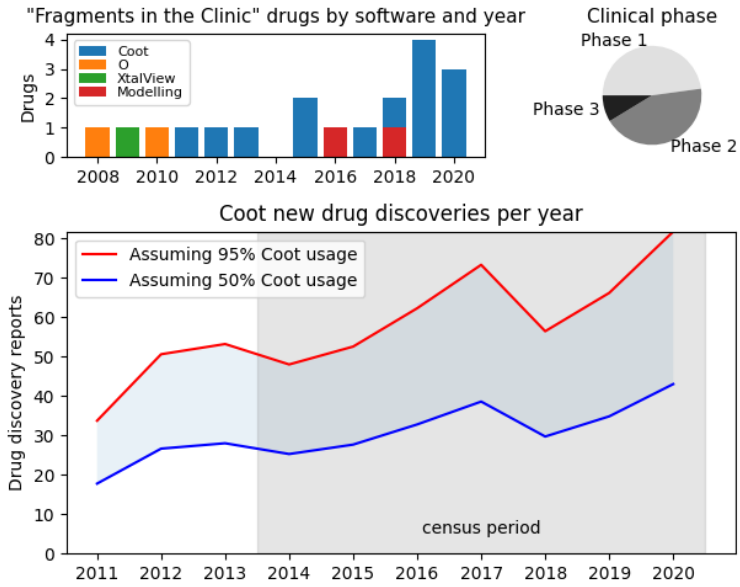
Structure-based drug design is an approach to the development of new treatments which employs the 3D structure of the drug complex to understand and develop better drugs. This typically involves data from X-ray crystallography or electron cryo-microscopy, processed using COOT and other YSBL-developed software. COOT plays an important role in allowing scientists to interpret their data to see the bound drug. York software therefore lies at the heart of early stages in the drug discovery process, particularly its enabling of successful Fragment-Based Drug Discovery (FBDD).

The number of possible drug-sized compounds that could interact with a target protein in the human body, and therefore support a pharmaceutical intervention, is so vast that computational methods are required to automate and streamline this first step in the development pipeline. Software-enabled FBDD addresses this with the targeted screening of a reduced subset of small-molecule fragments for weak binding with proteins. Combining weak binding fragments using structural information from crystallography can then produce effective drugs.

For instance, Carmot Therapeutics is one user of COOT software, and it produces a review of FBDD across the industry in its "*Fragments in the Clinic 2020*" report [5.2]. This shows that FBDD has recently generated four approved drugs and that there are currently 43 compounds in clinical trials from more than 30 manufacturers. Just one of those drugs (*Erdafitinib*, from Astex, a cancer treatment targeting tumor growth) is estimated to generate revenue of USD1.2 billion over five years [5.5]. We have traced the literature for each of those drugs in clinical trials and could identify the modelling software that was used in 23 of the new drug candidates [5.3]. The COOT software was cited in 17 of the 23 drugs for fragment fitting in the 'discovery paper', its supplement, or a linked methods paper. COOT supported the development of new drugs in clinical trials from manufacturers including Abbott, Amgen, Astex, AstraZeneca, Genentech, Lilly, Pfizer, Novartis, and Vernalis. Based on the median costs of clinical trials reported in *Nature Review Drug Discovery* (2017) this would represent an investment of ~USD150,000,000.

Other alternative softwares to COOT were mentioned in only four clinical trials cases, all of which date to before 2011. Two remaining cases relied on *in silico* modelling, although in one of these cases the protein structure itself was determined using COOT. This demonstrates how COOT has become the pre-eminent software tool of choice for FBDD in the pharmaceutical industry globally.

COOT is used by a diverse range of users and we have evaluated the wider scale of its global impact on the creation of new drugs by searching for 'discovery of' in the primary citation of the worldwide PDB (Protein Data Bank) entries (2011-2020), or in papers from *Scopus; pre-August 2013 entries are still relevant to this REF period as the drug discovery process continues over many years*. Based on PDB searches and citations, we identified 891 papers, from which we took a random subsample of 161, focusing in particular on papers with authors from the commercial sector. Of these 110 were drug discovery papers with commercial authors that directly used COOT as a key part of the drug development pipeline. Extrapolating these results to the full global literature indicates that between 2014 and 2020 there have been between **242** and **1029** commercial drug discovery papers involving COOT software based on information in the PDB. Applying the same type of analysis to papers included in Scopus leads to even higher estimates [5.4]. COOT-enabled FBDD drugs currently in clinical trials [based on 5.2] are shown in the figure below (*upper*) along with new drug discovery reports that cite COOT over the last decade (*lower*). [We show estimated drug discovery reports using the 95% usage data indicated by Scopus, and also, for reference, the number if we assumed a highly conservative COOT usage of only 50%].



Reading from this plot we are able to show that COOT became the method of choice in industry around 2010, with the period 2017-2020 showing particularly significant levels of growth in drug discoveries involving COOT usage.

We have undertaken consultations with a range of industrial and third sector users, who have highlighted the significant benefits of COOT in accelerating their drug discovery projects. One key comment is that its state-of-the-art fitting and validation tools can be used interactively *or* scripted for automation in high-throughput fragment screening [5.5].

York software in the context of COVID-19

High-throughput fragment screening for drug discovery has become possible through developments in data collection (*e.g.* crystal changing robotics and *in situ* screening of crystallisation drops) and structure determination (structure solution and refinement; finding and placing the bound ligand). The COOT software has played a crucial role in the acceleration of this process and in some cases the automation of the finding and placing step. The urgent need for screening for new approaches to the treatment of COVID-19 has seen the power of recent COOT developments come to the fore. This is exemplified by three projects which rapidly applied COOT and FBDD for the identification of potential new therapeutics [5.6 - 5.8]. These screened 1250, 5953 and 2683 compounds, respectively, for binding against SARS-COV-2 proteins, and identified 74, 37, and 234 ligand binding structures. The FBDD pipeline will now combine these hits to provide the starting point for identifying therapeutics to improve the treatment of COVID-19 in the post-pandemic phase. In parallel, the PRIVATEER software was used in the structural work (Pallesen, J. *et al.,* PNAS, 2020, Walls, A.C., *et al.* Cell, 2020) that led to the inception of the two mRNA vaccines that have received approval in the UK (Pfizer/BioNTech and Moderna Inc.), and several ultra-potent monoclonal antibodies against SARS-CoV-2.

Wider use of YSBL software and data innovations

The use of COOT software is not limited to drugs and therapeutics, and it finds application in several other areas of the chemical industries. For example Danisco and Goodyear reported the development of biological alternatives to isoprene in the production of rubbers and polymers, underpinned by COOT-generated data (see patent WO-2010124146-A2). The YSBL-developed CLIPPER libraries of software for crystallographic computation are now incorporated into other widely used structural and chemical software packages; for example the Isolde visualisation and model building software, an extension to the ChimeraX software, which is used in the rapidly expanding area of CryoEM research [5.9]. A wider search of patent databases shows COOT software being cited in 278 patents, as of end 2020 [5.10].

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

5.1. UK Bioindustry Association press release: https://www.bioindustry.org/news-listing/new-report-reveals-uk-has-the-strongest-rd-pipeline-in-europe-and-the-uk-sector-is-maturing.html

5.2. 'Fragments in the Clinic 2020', Dan Erlanson, co-founder of Carmot Therapeutics: https://practicalfragments.blogspot.com/2020/03/fragments-in-clinic-2020-edition.html

5.3. Spreadsheet of Erlanson's drug discovery list annotated with the associated PDB entries, papers and software used.

5.4. Text document describing the literature survey on which our estimates for the number of drug discovery papers are based. Two .csv files of supporting data are also provided with the submission.

5.5. Document containing material and testimony from interviews with commercial and third sector users of Coot, collected by "Fresh Perspectiv":

5.6. Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C.D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P. and Brandão-Neto, J., 2020. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nature communications*, *11*(1), pp.1-11. doi.org/10.1038/s41467-020-18709-w

5.7. Guenther, S., Reinke, P.Y., Fernandez-Garcia, Y., Lieske, J., Lane, T.J., Ginn, H., Koua, F., Ehrt, C., Ewert, W., Oberthuer, D. and Yefanov, O., 2020. Massive X-ray screening reveals two allosteric drug binding sites of SARS-CoV-2 main protease. *bioRxiv*. doi.org/10.1101/2020.11.12.378422

5.8. Schuller, M., Correy, G.J., Gahbauer, S., Fearon, D., Wu, T., Díaz, R.E., Young, I.D., Martins, L.C., Smith, D.H., Schulze-Gahmen, U. and Owens, T.W., 2020. Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. *bioRxiv*. doi.org/10.1101/2020.11.24.393405

5.9. Croll, T.I., 2018. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D: Structural Biology*, *74*(6), pp.519-530. doi.org/10.1107/S2059798318002425

5.10. Document containing patent citation counts for the two main COOT papers, provided by the Dimensions search engine.