



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Challenges of connecting chemistry to pharmacology: perspectives from curating the IUPHAR/BPS Guide to PHARMACOLOGY

### Citation for published version:

Southan, C, Sharman, J, Faccenda, E, Pawson, A, Harding, S & Davies, J 2018, 'Challenges of connecting chemistry to pharmacology: perspectives from curating the IUPHAR/BPS Guide to PHARMACOLOGY', *ACS Omega*. <https://doi.org/10.1021/acsomega.8b00884>

### Digital Object Identifier (DOI):

[10.1021/acsomega.8b00884](https://doi.org/10.1021/acsomega.8b00884)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Publisher's PDF, also known as Version of record

### Published In:

ACS Omega

### Publisher Rights Statement:

This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.

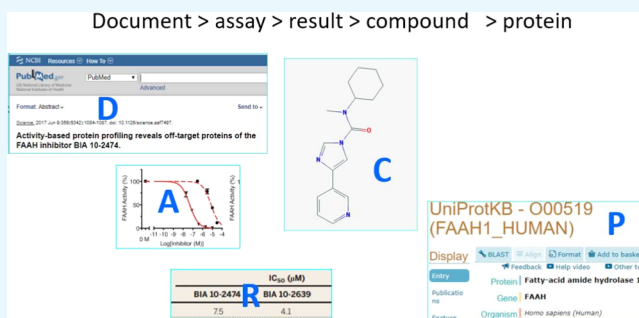


# Challenges of Connecting Chemistry to Pharmacology: Perspectives from Curating the IUPHAR/BPS Guide to PHARMACOLOGY

Christopher Southan,\*<sup>1</sup> Joanna L. Sharman, Elena Faccenda, Adam J. Pawson, Simon D. Harding, and Jamie A. Davies

IUPHAR/BPS Guide to PHARMACOLOGY, Centre for Discovery Brain Sciences, Deanery of Biomedical Sciences, University of Edinburgh EH8 9XD, Edinburgh, U.K.

**ABSTRACT:** Connecting chemistry to pharmacology has been an objective of Guide to PHARMACOLOGY (GtoPdb) and its precursor the International Union of Basic and Clinical Pharmacology Database (IUPHAR-DB) since 2003. This has been achieved by populating our database with expert-curated relationships between documents, assays, quantitative results, chemical structures, their locations within the documents, and the protein targets in the assays (D-A-R-C-P). A wide range of challenges associated with this are described in this perspective, using illustrative examples from GtoPdb entries. Our selection process begins with judgments of pharmacological relevance and scientific quality. Even though we have a stringent focus for our small-data extraction, we note that assessing the quality of papers has become more difficult over the last 15 years. We discuss ambiguity issues with the resolution of authors' descriptions of A-R-C-P entities to standardized identifiers. We also describe developments that have made this somewhat easier over the same period both in the publication ecosystem and recent enhancements of our internal processes. This perspective concludes with a look at challenges for the future, including the wider capture of mechanistic nuances and possible impacts of text mining on automated entity extraction.



## 1. INTRODUCTION

Connecting chemistry to pharmacology (c2p) can be understood intuitively as the effects of exogenous substances on the systems of living organisms, generally associated with exploring a human therapeutic application. However, it is necessary to limit our coverage of this extensive topic in this perspective. First, as implied from the title, this will focus on the c2p subset encompassed in the International Union of Basic and Clinical Pharmacology/British Pharmacological Society (IUPHAR/BPS) Guide to PHARMACOLOGY (GtoPdb, <http://www.guidetopharmacology.org>). This includes the addition of the new IUPHAR Guide to IMMUNOPHARMACOLOGY (<http://www.guidetoimmunopharmacology.org>). Second, it will concern mostly small molecules but also extends to moderate-size peptides and therapeutic nucleotides. Third, while it is not the only provenance option, we will restrict ourselves to scientific publications for the primary reports of pharmacology. Fourth, the focus is on effects that can be plausibly explained based on experimental results as a molecular mechanism of action (moma). The area to cover thus becomes more circumscribed as approximating to the document corpus covering pharmacology (i.e., largely journal papers and patents). We will also cover aspects of drug discovery and chemical biology, since these domains are not usefully separable from pharmacology. The same argument applies to metabolomics, but we leave this domain mainly to other resources such as the Human Metabolome Database.<sup>1</sup>

GtoPdb is a moderate-scale knowledgebase of c2p relationships, curated manually since 2003 from the pharmacological literature. Detailed content statistics and feature descriptions are available in our Nucleic Acids Research (NAR) Database issue article<sup>2</sup> and via our website and blog. GtoPdb has accumulated a matrix of 1700 targets and 15 500 activity values for 7000 ligands, guided by expert members of the IUPHAR Committee on Receptor Nomenclature and Drug Classification (NC-IUPHAR) and its 96 subcommittees. This de facto “super-curator network” encompasses over 500 individual scientists (<http://www.guidetopharmacology.org/nciuphar.jsp>).

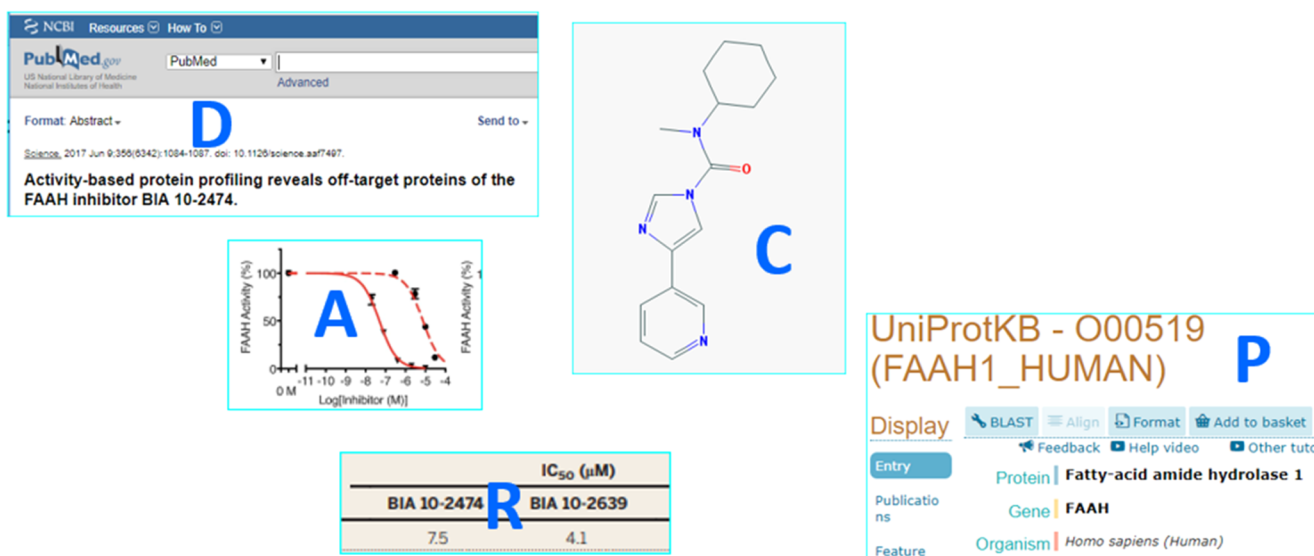
Our aim is to provide quantitative data on compounds suitable to use in the laboratory as recommended by experts. Consequently, GtoPdb has a high level of selectivity (including being mainly focused on human data). Our quality selection criteria were described for our first publication back in 2009.<sup>3</sup> We use several indicators to demonstrate the value impact of this approach across our broad user community. These include our user counts (via Google Analytics) of 19 000 per month from 200 countries, web services accesses, data downloads, good citation rates for all six of our NAR Database issue papers, cross-linking from over 20 other sources, and our 2016 inclusion in the

Received: May 2, 2018

Accepted: July 12, 2018

Published: July 31, 2018

## Document &gt; assay &gt; result &gt; compound &gt; protein



**Figure 1.** Example of a D-A-R-C-P relationship chain. Taken from <http://www.guidetopharmacology.org/GRAC/LigandDisplayForward?ligandId=9001>.

ELIXIR UK Node resources for Human Health and Disease (<http://www.elixiruknode.org/human-health-and-disease/>).

## 2. DESCRIBING CHEMISTRY TO PHARMACOLOGY (C2P) CONNECTIONS

**2.1. D-A-R-C-P Concept.** We can conceptualize c2p as core relationships (i.e., chain of links) between a document, an assay, its quantitative result, a chemical structure, and a protein target.<sup>4</sup> A useful shorthand for this is D-A-R-C-P (note the order is not crucial but will be adhered to here for simplicity). A topical example is illustrated in Figure 1.

In this case “C”, as BIA 10-2474, is the code number of the BIAL Portugal, Phase I compound that caused a fatality during a French clinical trial in January 2016.<sup>5</sup> The following c2p relationships can be extracted:

- “D” is a 2017 paper entitled “Activity-based protein profiling reveals off-target proteins of the FAAH inhibitor BIA 10-2474” (although not from BIAL). Thus, “D” can be specified as PMID 28596366, or by its DOI: 10.1126/science.aaf7497.
- From this document, we can extract “A” as an IC<sub>50</sub> measurement on purified protein in vitro, which records a (low) potency result “R” of 7500 nM. R can also be expressed as 7.5 μM or a pIC<sub>50</sub> of 5.12.
- We can specify “C” by an IUPAC name *N*-cyclohexyl-*N*-methyl-4-(1-oxidopyridin-1-ium-3-yl)imidazole-1-carboxamide. Alternatively, we could use the PubChem Compound ID CID 66554294,<sup>5</sup> the GtoPdb Ligand ID (LID) 9001, or an InChIKey DWCWWJONKWHPPDD-UHFFFAOYSA-N (full InChI strings or SMILES can also be used).
- The entity “P” could use the HGNC-approved gene name for the enzyme,<sup>6</sup> as “fatty acid amide hydrolase”, its symbol, *FAAH*, or the Swiss-Prot human protein accession O00519.

While such D-A-R-C-P mappings may seem highly specific, there are other useful relationships that can be extracted. For

example, while the publication is “about” BIA 10-2474 and the inhibition of the FAAH enzyme, it also includes the putative off-targets and extrapolates to the paralogue FAAH2 (by homology). We can also determine that CID 66554294 has 1199 nearest neighbors by chemical similarity within PubChem, thereby inferring that FAAH inhibitors are (or at least have been) an active area of research. Typically, in structure–activity relationship (SAR) papers, there are multiple “R-C”s for the same “D-A” and P. Thus, within PMID 28596366, we can identify two additional C’s as the metabolite BIA 10-2639 (CID 66554294) and a reference inhibitor PF-04457845 (CID 24771824), where R is 4100 and 10 nM, respectively.

Figure 2 illustrates additional c2p inferences from powerful features of the PubMed/NCBI Entrez system.<sup>7</sup> From the available “D-D” links, we can find useful further references, including a primary reference for the first declaration of in vitro activity measurements against PF-04457845.

**2.2. c2p Curation in GtoPdb.** GtoPdb utilizes expert curation to identify, extract, and convert unstructured knowledge from the published article (as text, chemical images and result tables) into structured database entries. This can be illustrated for the example above by looking at the table of entries GtoPdb has selected for FAAH inhibitors shown in Figure 3.

This includes the three D-A-R-C-Ps already mentioned for BIA 10-2474, BIA 10-2639, and PF-04457845, with the latter having two D’s and two R’s indicated as a range. In the FAAH table, we have 15 unique C’s (as ligands), 23 R’s (as affinities), and 13 D’s (as references). We also have four instances of rat data. In these cases, D-A will be the same, R values will be different, and P will be P97612 for the rat enzyme. If we inspect the analogous records for human FAAH2 (i.e., where P = Q6GMR7, GtoPdb TID 1401), we can find inhibition data for four of the compounds. Importantly, in this case, there is no rodent orthologue for FAAH2.

Over the years, we have extended our data model to support additional relationships. One of these was a sixth attribute in the



**Figure 2.** Relationships within the Entrez system available as links from the right-hand facets of a PubMed entry. “Similar articles” is essentially a heuristic clustering of “aboutness”, while “cited by” provides forward connectivity within the same knowledge domain. “Related information” links to NCBI database cross-references including PubChem Substance (SID), Compound (CID), and BioAssay, MeSH keyword matches, and protein structures in the Protein Data Bank (PDB). In this case, the three “PubChem substance” links have entered the system via the PubChem submissions from GtoPdb (see section below on PubChem links). The “MeSH keyword” look-up brings back three CIDs for BIA 10-2474, PF-04457845, and the less relevant urea (n.b. MeSH annotators did not select the metabolite BIA 10-2639 as a keyword but could do if this becomes a main theme of a future paper).

link, “L”, as a reference to the explicit location(s) of C within the document D. Our shorthand was thus extended to D-A-R-C-L-P. The utility is that from a series of compounds specified in a typical medicinal chemistry paper, we specify the intradocument named leads, most often as listed compound names from an analogue series and/or the code number of the declared lead compound. Users can thus quickly see both what the paper is about and how C has been provenanced in terms of curatorial choice. This innovation was helpful as we began to include patents as D’s (which often extend into 100s of IC<sub>50</sub>s), as a consequence of expanded patent-extraction content in PubChem.<sup>5</sup> An example from Figure 3 is ligand ID 9077, showing the location as “example 13 [WO2009109743]” (i.e., example 13 is L for its exact location in the document), which was selected as a representative potent inhibitor from a patent filed by the company Vernalis.






























We informally refer to the citations we associate with our ligands as primary, secondary, or tertiary publications. “Primary citation” has already been described above as A-R-C-L-P where the data are obtained with purified protein in vitro. We use the term “secondary citation”, where the publication D is about in vivo testing in animal models and typically cites back to a primary paper. “Tertiary citation” is applied to clinical reports

that typically cite both primary and secondary studies. These divisions are somewhat arbitrary and not without limitations, but in practice, we find that this classification offers users a small set of core linked references, anchored to the same “C-P”, as starting points for data mining and further exploration of the literature.

Another important aspect of citations associated with our curated references for ligands is the dissemination of these in PubChem. GtoPdb has been a submitter to PubChem since 2008, but in the last few years, as a consequence of the reference curation expansion described above, we are now one of the major expert annotation sources of “C-D” (see Table 1 in ref 5). In addition, essentially via the processes above, we are also one of the most stringent in terms of key compounds. All of these are in D-A-R-C-P chains and many are full D-A-R-C-L-P. The statistics are shown in Table 1.

### 3. CHALLENGES OF c2p CURATION

One of the challenges for GtoPdb is to cover c2p with modest biocurator resources, supported by the work of NC-IUPHAR subcommittees. We address this by exercising judgment on what to include and, equally importantly, what to leave out. We also may not pick up early development leads or even new phase 1

Inhibitors							
Key to terms and symbols		View all chemical structures			Click column headers to sort		
Ligand		Sp.	Action	Affinity	Units	Reference	
JNJ40355003	 	Hs	Inhibition	8.9	pIC <sub>50</sub>	8	▼
JZL195	 	Hs	Inhibition	8.7	pIC <sub>50</sub>	13	▼
PF-04457845		Hs	Inhibition	8.0 – 9.0	pIC <sub>50</sub>	7,16	▲
	pIC <sub>50</sub> 8.0 – 9.0 (IC <sub>50</sub> 1×10 <sup>-8</sup> – 1×10 <sup>-9</sup> M) [7,16] Description: Inhibition of human FAAH <i>in vitro</i> .						
compound 19y [Kiss <i>et al.</i> , 2011]		Rn	Inhibition	8.2	pIC <sub>50</sub>	12	▼
(S)-ARN2508	  	Hs	Inhibition	8.0	pIC <sub>50</sub>	14	▼
(R)-ARN2508	  	Hs	Inhibition	8.0	pIC <sub>50</sub>	14	▼
JNJ1661010	  	Hs	Inhibition	7.8	pIC <sub>50</sub>	9	▼
JNJ40355003		Rn	Inhibition	7.5	pIC <sub>50</sub>	8	▼
example 13 [WO2009109743]		Hs	Inhibition	7.4	pIC <sub>50</sub>	15	▼
OL135	 	Hs	Inhibition	7.4	pIC <sub>50</sub>	18	▼
BIA 10-2474		Hs	Inhibition	7.2 – 7.3	pIC <sub>50</sub>	16	▼
BIA 10-2639		Hs	Inhibition	7.2 – 7.3	pIC <sub>50</sub>	16	▼
JNJ-42165279	 	Hs	Inhibition	7.1	pIC <sub>50</sub>	10	▼
PF750		Hs	Inhibition	6.3 – 7.8	pIC <sub>50</sub>	1	▼
BIA 10-2474		Rn	Inhibition	~6.7	pIC <sub>50</sub>	11	▼
URB597	 	Hs	Inhibition	6.3 – 7.0	pIC <sub>50</sub>	18	▼
PF3845		Hs	Inhibition	6.6	pIC <sub>50</sub>	2	▼
JNJ-42165279		Rn	Inhibition	6.5	pIC <sub>50</sub>	10	▼

**Figure 3.** GtoPdb entry for FAAH (GtoPdb Target ID (TID) 1440) with inhibitors mapped to it from the database release 2018.1. The record for PF-04457845 is expanded to show the activity and references. Descriptions of the icons and rows are given in the GtoPdb Help documentation and FAQs.

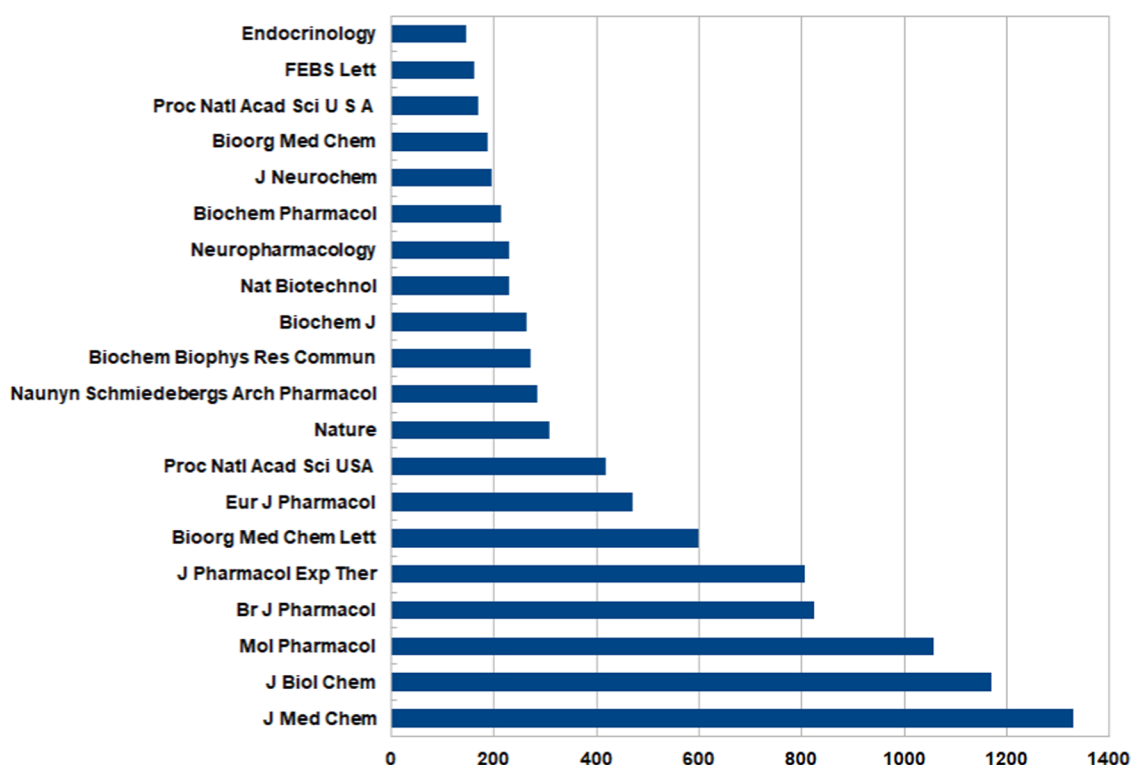
**Table 1. Statistics Associated with GtoPdb Relationship Extractions**

unique PubMed IDs linked to target–ligand pairs (C-P)	6466
unique PubMed IDs for primary D with quantitative “A-R”	6078
primary D with quantitative A-R, where C has a CID	5246
references linked to our substance identifiers (SIDs) in PubChem	9253
average compounds-per-document for A-R	1.1

compounds for popular targets toward the top end of our ligand distribution (e.g., the dopamine D<sub>2</sub> receptor TID 215 with 105 ligands in the download list). However, we try to capture all approved drugs and phase III candidates. As further examples of selectivity judgments, our BACE1 entry (TID 2330) has 21 inhibitors listed compared to 8545 for the ChEMBL (release 24) entry. For another historically popular target, F2 (thrombin), the GtoPdb/ChEMBL mapped compounds ratio is 10:8264. However, since we are not constrained by coverage mandates, we can introduce committee-recommended positive biases. Examples here include the recent crop of G-protein-coupled receptor (GPCR)-biased agonists and allosteric modulators, where we have curated those with PDB structures of ligand binding pockets. In addition, we have introduced a new tool, SynPharm, to allow synthetic biologists to exploit PDB ligand binding sequences.<sup>8</sup> We also use our judgment in the opposite

direction (i.e., lowering the bar) in that we pick up low-potency compounds directed against novel targets or orphan GPCRs while maintaining a potency cutoff of ~10 μM. This is because these papers represent new pharmacological starting points and thus extend the druggable proteome coverage. More potent compounds may eventually be curated against the same target, but we generally do not purge data-supported first-generation ligands.

A second crucial challenge for c2p is to ensure (as far as possible) that the reported experimental results we choose to curate are correct. This has become harder as GtoPdb domain coverage expands into new areas (especially enzymes and other nonreceptor or channel proteins, which were our original focus prior to 2012), journals have proliferated, and there is a lower proportion of NC-IUPHAR subcommittee coverage within newer areas. These combined factors have made quality judgments more difficult. This challenge is now compounded (and thrust into the foreground) by the “reproducibility crisis” affecting the experimental biosciences in general and rising rates of translational failure for pharmacology in particular.<sup>9</sup> Major concerns began surfacing in 2010 with a report indicating that mislabeled or misidentified cell lines had affected thousands of papers.<sup>10</sup> This was followed in 2011 by the company Bayer declaring that their in-house experimental findings failed to match up with 65% of published target



**Figure 4.** Distribution of papers curated for ligand interactions in GtoPdb (release 2018.1). Numbers of papers are shown on the horizontal axis. The vertical axis of journal titles shows the top-20 journals from a total of 920.

validation claims,<sup>11</sup> and in 2013 by a study pointing out dispensing errors in assays.<sup>12</sup> By 2015, poor antibody specificity was also coming under the spotlight.<sup>13</sup> At about the same time, concerns were being raised about the inappropriate use of statistics and other aspects of experimental design, which prompted revisions of the author guidelines for the Br. J. Pharmacol.<sup>14</sup> These cautionary tales were compounded by a 2017 report that 29% of over 8500 vendor drug compounds failed purity quality control.<sup>15</sup> As if all this was not enough (while not directly an experimental reproducibility issue per se), the increasingly documented shortcomings of rodent models for human diseases (e.g., those most often used in our secondary citations) are being blamed for unsustainable rates of translational failure across all therapeutic areas.<sup>16</sup>

So, what can GtoPdb do in the face of all this? The first thing is to maintain curatorial vigilance and stringency, backed up by both NC-IUPHAR subcommittee oversight and direct user feedback. This is helped by NC-IUPHAR's familiarity with the global research teams working in these areas over many years. They can thus provide quality judgments of both the literature and our database records curated from it. An example of this is the subcommittee that recommends criteria for the interaction reported for an orphan receptor with new cognate ligand(s) and continues to update confirmed pairings.<sup>17</sup> Also, our database team (past and present) are scientists who have published in their own right and are also members of the International Society for Biocuration.<sup>18</sup>

We have taken other strategic options that, while certainly not eliminating the problems, go some way to ameliorating them. One of these is that we now enter the curatorial cycle more often via publications that give explicit support to primary experimental reports (i.e., we increasingly initially pick up ligands and/or targets via a secondary publication, meaning they

are supported by at least two papers). In the interests of reproducibility, we also try to find at least partially overlapping confirmations from independent teams. We also routinely exploit the PubMed features of both similar articles and cited by (Figure 2). This gives us a convenient "360° walk" of any publication as an adjunct QC. The former should pick up publications on the same target and the latter may provide corroboration of that activity modulation approach. However, perhaps the dominant factor is that we do not aim for total c2p coverage. This means we can be highly selective for small-data coverage of pharmacology.

Two specific examples illustrate the importance of our c2p quality judgments. The first concerns article retractions where there is evidence of research misconduct or error.<sup>19</sup> As a cross-check, we have recently selected the 5801 retracted entries in PubMed and checked which sources had added the 226 linked SIDs (although this seems a reassuringly low proportion of the total PubMed linkage count of 260 754). Of these, 225 turned out to be derived from the automated IBM pipeline for patent extraction that also operates on PubMed abstract texts. However, the 226th was our own entry for lysophosphatidylcholine (LID 2508). It turns out that this had been initially associated with a deorphanization report suggesting GPR4 activation by lysophosphatidylcholine and sphingosylphosphorylcholine (LID 4032) but which was later retracted.<sup>20</sup> Rather than simply breaking the link that would lead to a loss of contextual information, we have explained the retraction in the comment fields for both ligands. While such retractions will only pick up a small proportion of irreproducible c2p, they have been increasing due to enhanced detection methods, so we will continue to regularly intersect our reference lists.

The second cautionary example is an exclusion judgment. During the updating of dipeptidyl peptidase-4 (DPP4)

inhibitors (TID 1612), we picked up a reported low-potency interaction between this enzyme as a successful diabetes target and atorvastatin, recorded in several databases.<sup>21</sup> The original paper declares an  $IC_{50}$  of 175  $\mu M$  and  $K_i$  of 58  $\mu M$  for atorvastatin against pig DPP4 in vitro. However, this exceeds our usual inclusion threshold of  $\sim 10 \mu M$  (although not applied as an absolute rule) by nearly an order of magnitude, even for off-target cross-reactivity. By comparison, we had curated the human DPP4 drug omarigliptin (LID 8402) with an  $IC_{50}$  of 1.6 nM and  $K_i$  of 0.8 nM, and atorvastatin (LID 2949) as having an 8 nM  $IC_{50}$  against human hydroxymethylglutaryl-CoA reductase. To be clear, we are not questioning the authenticity of the atorvastatin versus DPP4 result from ref 21 and we do take other primary ligand reports from this journal. Notwithstanding, we not only remain unconvinced as to the in vivo pharmacological relevance but also note the absence of any subsequent in vitro confirmatory reports. While our own judgment was therefore to exclude it, we noted that this interaction had been captured by other sources. The first of these was in DrugBank,<sup>22</sup> where the interaction (although not classified as pharmacologically significant) is included, with human DPP4 listed as one of the targets of atorvastatin (DB01076) but without activity values. This contrasts with ChEMBL<sup>23</sup> in which the same paper was extracted but the database record maps the atorvastatin  $IC_{50}$  against the correct pig enzyme (ChEMBL3813). The ChEMBL entry was then transitively propagated to both BindingDB (atorvastatin as BDB 22164) and PubChem BioAssay (as AID 313843).

**3.1. Challenges Associated with D.** The challenge of document curation (D) is that of scale and selectivity. Estimating the total extractable D-A-R-C-P from the entire document corpus relevant to pharmacology, drug discovery, and chemical biology is difficult, to say the least. However, one commercial source of such mappings, Excelra (previously GVK BIO), has declared 1.34 million compounds from 112 000 curated papers in 2016, thus recording an average of  $\sim 12$  C's per paper.<sup>24</sup> We can also compare with public resources, such as ChEMBL,<sup>23</sup> where release 24 declares 1.25 million structures extracted from just under 70 000 documents (including 62 205 PubMed IDs according to European PubMed Central (EPMC) indexing). This gives an average of  $\sim 18$  C's per paper in this case. The figures from BindingDB<sup>25</sup> (for papers not overlapping with ChEMBL) are 3385 PubMed IDs and 43 397 compounds, with a slightly lower average of  $\sim 14$  C's per paper (Gilson, personal communication).

Compared to these other sources, GtoPdb C's per paper average of 1:1 reflects our typical selection of a single lead from each primary paper (Table 1). One way to monitor our selectivity is by the distribution of journal papers we curate associated with ligands (Figure 4).

The distribution is long-tailed (i.e., 43% are singleton journals), but we can detect the splits between our empirical classifications. Thus, *J. Med. Chem.* tops the list of primary reports. As we have defined them above, secondary papers are widely distributed with *Br. J. Pharmacol.* being a good example (although it includes some primary reports). As expected, tertiary reports are fewer since only a minority of lead compounds from the primary and secondary papers make it through to clinical publications such as *N. Eng. J. Med.* or *Br. J. Clin. Pharm.*

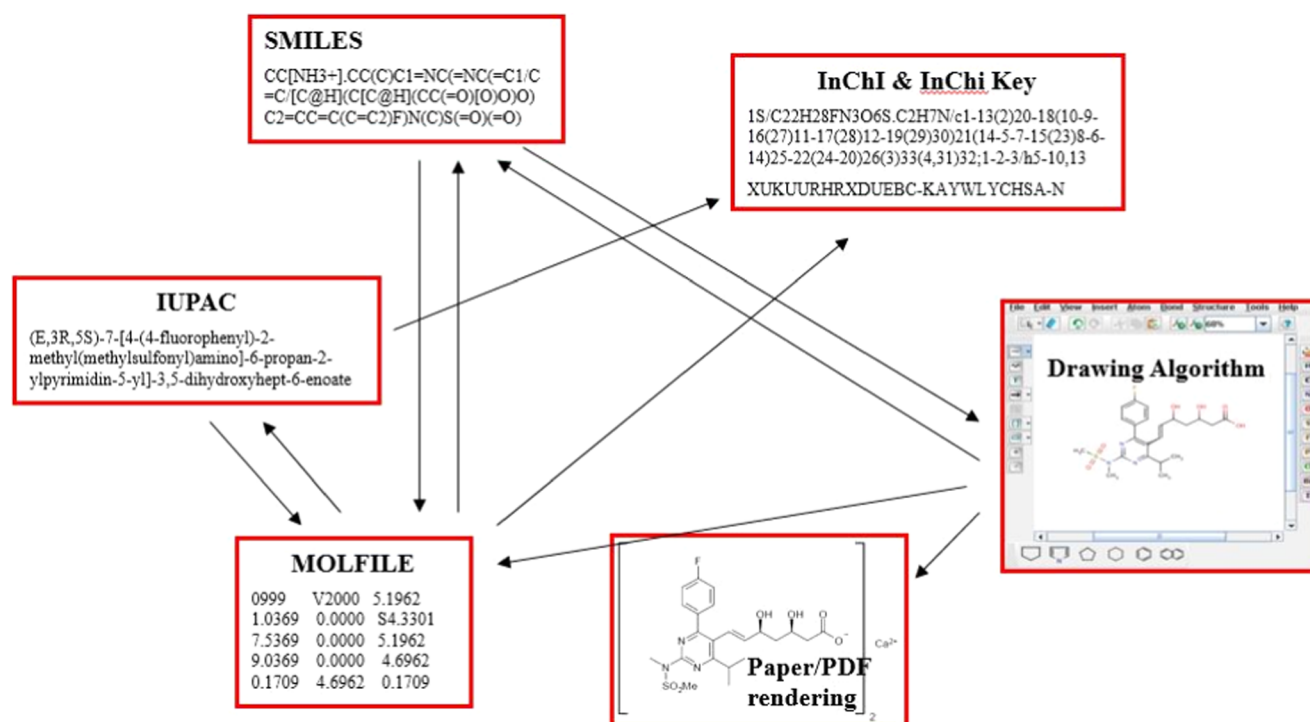
Other negative trends we have become aware of (besides the reproducibility crisis outlined above) include proliferation (i.e., c2p being spread over 920 journal layouts with which we must

individually grapple with for curation), "salami-slicing" (i.e., data from essentially the same study being divided between smaller publishable units), and "me-too" papers. These ostensibly disclose new data but for low-potency compounds unlikely to ever be optimized. Another disappointing aspect is the relatively slow penetration of Gold Open Access (GOA) into the medicinal chemistry literature. Advantages of GOA include that full-text HTML facilitates faster and more accurate curation (e.g., reducing formatting errors associated with pasting IUPAC names out of PDFs). In addition, both PubMed Central (PMC) and EPMC not only index full text for searching but also for different types of automated entity indexing (e.g., gene names). As a domain, pharmacology has done somewhat better in open access (OA) statistics than medicinal chemistry. For example, the *Br. J. Pharmacol.* achieved 77% PubMed Central (PMC) free full-text conversion for their 2001 PubMed IDs over the last 5 years (i.e., with a 6-month embargo period). The recently launched sister journal, *Pharmacol. Res. Perspect.*, was conceived as GOA from the outset and thus has a PMC full-house for its 356 PubMed entries.

**3.2. Challenges Associated with A.** We made an early decision to not add assay details to our database records because we defer to the papers that we select to describe A (for those subsets of R that we also select). There were two main reasons for this. First, we had neither the curatorial capacity to add detailed assay descriptions nor (until the BioAssay Ontology becomes widely adopted) would we have a classification for them.<sup>26</sup> Second, while we do add brief notes for unusual or new types of assays, it is difficult to make short summaries that experimentalists could use without missing essential details, which means they would have to consult the paper anyway. In addition, even in the papers, descriptions often turn out to be incomplete (e.g., missing precise buffer conditions). We feel journals could do more here to mandate reproducible descriptions since guidelines have been made available for many years. These include Standards for Reporting Enzyme Data<sup>27</sup> and Minimum Information about a Bioactive Entity.<sup>28</sup> However, it has to be accepted that across the major target classes (GPCRs and ion channels in particular), assays are becoming more complex and thus making editorial oversight more difficult.<sup>29</sup>

**3.3. Challenges Associated with R.** We curate R as a standard activity modulation parameter, where  $IC_{50}$ ,  $EC_{50}$ ,  $K_i$ , and  $K_d$  are in the majority. We typically report these as specified by authors, but we normalize concentrations to nanomolars and then log these to pAct. We do not always accept the primacy of what is in the paper but will correct in our database records that which we judge as clearly erroneous. In practice, such corrections are rare since, if we find that the frequency of probable mistakes in any paper is high (due to questionable quality or lack of editorial stringency), it is not selected for curation. What we occasionally do fix are obvious unit mix-ups between micromolars and nanomolars. A more common issue is the quoting of significant figures way beyond those appropriate to the error ranges of the assay. An example is SID 103716988 from PubChem BioAssay (AID 566893) with a BACE1  $IC_{50}$  value of 1.38995  $\mu M$ . Since the error of such protease substrate turnover assays typically exceeds 10%, we would have rounded this to three significant figures (but we did not select this particular paper).

We also come across activity units that we may not typically include but will mention in a comments field. For example, covalent enzyme inhibition can be expressed as K (inact) in a



**Figure 5.** Six principal types of chemical representation or routes of interconversion encountered and/or used during the curation of structures from papers. Most cheminformatics tool-kits can execute the interconversions indicated by the arrows and major chemistry databases will also precompute links between them. However, the round-tripping may not be perfect. Note also that the InChIKey cannot be converted to a structure but has key utilities, including as a look-up string in Google.

paper, but we will use the IC<sub>50</sub> at fixed preincubation time.<sup>30</sup> If bioactivity is only reported as percentage inhibition at fixed concentrations, then we take the absence of dose response data to indicate that the work is more of a preliminary investigation. While we would not usually curate such paper, we make exceptions for what we judge as significant pharmacology from difficult assays. We also do not normally add reported error ranges to database records. This is because of lack of standardization in different authors' statistical treatment of variance ranges as well as often a lack of clarity as to whether the variation for technical or biological replicates was being measured. As part of our procedure, we inspect PubChem BioAssay<sup>31</sup> records (i.e., A-R) that may already be linked to C. We suspect few authors perform this cross-check (or at least do not mention it). Sometimes this indicates to us that their compound (or a close analogue) is in fact more potent against a target that was not cross-screened in their paper. In such cases, we mention this in curators notes.

**3.4. Challenges Associated with C.** Since curating the C in D-A-R-C-P is arguably the most critical link for c2p, we employ a series of curatorial steps. First and foremost, we must discern if the authors have unequivocally specified a defined structure for C that is either novel (as defined by its absence in GtoPdb or other databases) or already exists in some other provenanced source. We call this process a structure-to-structure mapping (s2s). Second, we resolve any names (semantic, common, nonproprietary, code names, and synonyms) against that structure. We term this name-to-structure (n2s), which includes Google checking. We may also have to resort to a third process of image-to-structure (i2s) if this is the only explicit description of C in the paper. We note that these processes are reciprocal and that users may need to perform them in different searching contexts (e.g., "what name does this

structure search retrieve from GtoPdb" is essentially a structure-to-name query). The standard specifications we use for chemical curation are shown below.

We have previously outlined reasons why we have chosen PubChem as our primary source of C.<sup>32</sup> Our s2s process is thus centered on mapping key structures in papers to Compound Identifiers (CIDs) and we subsequently submit our own records to PubChem as Substance Identifiers (SIDs). The GtoPdb release 2018.1 thus includes 6969 CIDs we have resolved by curation, but in 95 cases, we are the sole source. These are usually associated with novel structures but, as discussed below, some may be alternative representations of pre-existing structures. So why are our CID s2s mostly successfully mapped? The main reason is that PubChem has now expanded to 95 million CIDs from 598 sources, thereby (from the alternative databases of comparative size) becoming the de facto global chemistry and bioactivity hub of choice.<sup>33</sup> Our largest source type overlap within PubChem of 80% comes from patents, reflecting the fact that most structures in drug discovery reports have already surfaced in PubChem via automated patent extraction.<sup>34</sup>

The spectrum of equivocality and even frank errors we regularly encounter in resolving C during s2s operations is frustratingly broad and some of them have been reviewed elsewhere.<sup>35,36</sup> This can be largely attributed to the reluctance of journals to mandate the use of public database identifiers for chemical structures. By this omission, they lag several decades behind molecular biology and bioinformatics, where editors have long insisted on accession numbers for protein and nucleic acid sequences to be reciprocally linked to PubMed records (although even here, no journal actually had complete compliance when surveyed in 2006).<sup>37</sup> Since PubChem has had stable chemical identifiers for well over a decade, it is unclear



A	B	C	D	E	F	G	H	I	J	K	L
Compou	SMILES	BACE1 Ki	BACE1 Cel	CatD/BAC	Caco-2 pe	Caco-2 eff	rat Ab40 C	rat Ab40 c	Rat Plasm	rat brain/i	rat hepatic
3	<chem>C[C@@](CS(N1C)=O)(NC1=N)C2=</chem>	2.2	2.1	>10000	128	2.4	-85	-53	36	0.57	3.1
7	<chem>N=C(N1C)N[C@@](C)(CC1=O)C2=C(C=</chem>	1.7	11	21	153	2.4	-81	-61	33	2.9	
9	<chem>N=C(N[C@@](C1=C(C)C)C=C(C2=CC(C#</chem>	2.4	55	47	278	1.6	-82	-53	8	2.3	
12	<chem>N=C(N1)N(C)C(C1(C2=CC(NC(C3=NC=C</chem>	696		22							
13	<chem>N=C(N1)N(C)C(C1(C2=CC(NC(C3=NC=C</chem>	595		83							
14	<chem>N=C(N1)N(C)C(C1(C2=CC=C(C3=CC(C#</chem>	14000									
15	<chem>N=C(N1)N(C)C(C1(C2=CC(C3=CC(C#N</chem>	848									
16	<chem>N=C(N1)N(C)C(C1(C2=CC(C3=CC(C#N</chem>	130									
17	<chem>N=C(N[C@@](C(S1)=CC=C1C2=CC(C#N</chem>	88									
18	<chem>N=C(N[C@@](C1=CC(C2=CC(C#N)=CC</chem>	57									
19	<chem>N=C(N[C@@](C1=CC(C2=CC(C#N)=CC</chem>	350									
20	<chem>O=C(C1=NC=C(C(F)(F)F)C=C1)NC2=CC</chem>	30	162	234							
21	<chem>C[C@@](CC(N1C)=O)(NC1=N)C2=CC(f</chem>	15	26	>1000	25	8.5	-42	-26	38	0.28	23
22	<chem>C[C@@](CC(N1C)=O)(NC1=N)C2=CC(f</chem>	4.2	3	>1000	<1	>23	-4		2	0.67	35
23	<chem>C[C@@](CC(N1C)=O)(NC1=N)C2=CC(f</chem>	4	4.6	>10000	22	11	-53	-11	22	0.38	2.6
24	<chem>C[C@@](CC(N1C)=O)(NC1=N)C2=CC(f</chem>	0.8	2	>10000	<1	>16	-63	-19	30	0.34	8
25	<chem>C[C@@](CC(N1C)=O)(NC1=N)C2=CC(f</chem>	3.8	7	>10000	17	11	-61	0	43	0.45	4
26	<chem>C[C@@](CS(N1C)=O)(NC1=N)C2=</chem>	10	5.8	>1000	92	2.3	-35		29	0.52	
27	<chem>C[C@@](CS(N1C)=O)(NC1=N)C2=</chem>	5.1	3.4	>10000	72	3.8	-8		23	0.81	
28	<chem>C[C@@](CS(N1C)=O)(NC1=N)C2=</chem>	7.1	7.5	>1000	74	3.9	-14		16	0.49	14
29	<chem>C[C@@](CS(N1C)=O)(NC1=N)C2=</chem>	3	2	>10000	151	2.1	-17		0.5		44
30	<chem>C[C@@](CS(N1C)=O)(NC1=N)C2=</chem>	0.9	1	>10000	118	3.1	-93	-51	17	0.89	13
31	<chem>C[C@@](CS(N1C)=O)(NC1=N)C2=</chem>	1.1	0.7	>10000	89	2.9	-75	-31	25	0.6	11
32	<chem>C[C@@](CS(N1C)=O)(NC1=N)C2=</chem>	13	15	>1000	14	16	18		9	0.39	6

Figure 6. Supplementary data from ref 39 with verubecestat as “compound 3”.

why such journal requirements have not appeared. One exception is *Nat. Chem. Biol.*, where author-specified SID structures are submitted by the journal. The current metrics are 7949 SIDs from 2444 papers (but some are on-hold pending publication). We have cross-referenced 51 of these papers and linked 81 ligands. As a primary source journal, it lies just outside the top-20 in Figure 4, since this is less mainstream for pharmacology. However, the pre-established s2s is valuable and makes it usually straightforward to link up the other entities. As our largest primary source of compound extractions, *J. Med. Chem.* has moved some way toward s2s with the introduction in 2014 of author-specified SMILES strings as open CSV tables (Figure 5).<sup>38</sup> For example, during curation, we cross-checked the BACE1 inhibitor clinical compound for Alzheimer’s disease, verubecestat (LID 8699), against the data sheet in Figure 6.

The inclusion of SMILES and their associated bioactivities expedites GtoPdb curation. In the absence of this, in most other journals, we have to pull out IUPAC text and convert this to SMILES, InChI string, and InChIKey using either the OPSIN<sup>40</sup> or Chemicalize.org web tools.<sup>41</sup> However, we find that names used by authors may need editing to remove formatting quirks or typographical errors. In the absence of IUPAC names (e.g., Markush enumerations for an SAR series), we may have to resort to i2s (as SMILES) with OSRA.<sup>42</sup> Raw i2s outputs from this often need correcting but usually have a close enough similarity match from a PubChem search. For poor images, we resort to a chemical sketcher that outputs SMILES. We then search PubChem via either SMILES, InChI string, InChIKey, SDF, or MOL file (i.e., the standardized formats from Figure 5), where we can usually resolve s2s via a CID exact match. We also cross-check against what is in the paper to clarify the specification of either enantiomeric mixtures and/or the resolved stereoisomer forms as R/S and E/Z. Our last operation on the compound entries is analogous to what we do for the papers in PubMed: a 360° walk of the structural neighborhood within PubChem.

This includes the precomputed relationships with similar compounds (analogues), same connectivity (i.e., different isomers or tautomers of the same skeleton), and mixtures, which can have bioactivity data that may not be recorded for parent compounds. In terms of deciding correctness (including isomerism), our empirical judgment is if many PubChem submitters agree on the structure (i.e., for a clear majority of substances merged in the CID) and the names in multiple sources also match, minimum parsimony would indicate both s2s and n2s be correct.

Pharmacologically active peptides as C entities present different s2s and n2s problems to small molecules. They require options of representation as three-letter code or FASTA sequence strings. Endogenous, unmodified peptides specifically cleaved from precursors are usually specified in the UniProtKB feature lines. Many are also specified as SMILES in PubChem CIDs. Complications arise from author-declared names that may not be IUPAC standard and/or include post-translational modifications, such as N-acetylation or C-amidation. Exogenous, synthetic peptides have the same issues and, moreover, are often not found in sequence databases. Radioactively labeled analogues (for small molecules as well as peptides) also feature prominently in the pharmacological literature. However, sometimes the molecular position of the label is unspecified, so we must duplicate the structure record to point to distinct references where data are generated with the labeled compound.

One negative aspect that has not changed in the last decade is that approximately 40% of all declared drug development candidates (i.e., where basic pharmacology and therapeutic indications are partially outlined on company portfolio websites and press releases) still have no associated structure (Lloyd, personal communication). The unappreciated numerical scale of the problem is reflected in the total of 7855 small-molecule drug projects declared globally in 2017.<sup>43</sup> Thus, the 40% would represent over 3000 blinded compounds. What makes matters

worse is that many of these have entered their clinical phases, even up to and beyond phase II, without explicit n2s in the literature (so we make the judgment not to capture these in GtoPdb). It is likely most of these structures are already in PubChem, having been extracted from patents, but these C entries do not specify which are the full leads or development candidates (and the patents claiming the structure series are often filed before these were chosen). This problem was highlighted in 2012 in the context of stalled industry clinical candidates offered as repurposing proposals, but where n2s could not be found for many of the code numbers.<sup>44</sup> This pernicious and pervasive practice of c2p “blinding” (practiced by academic drug discovery operations as well as commercial ones) is based on the fear of competitive fast-following. Notwithstanding, such blinding remains the single biggest obstacle to capturing recent development compounds (i.e., by definition, the cream of the pharmacological crop). It should be noted that there is no statistical evidence that has established that delaying open publication (or not publishing at all) reduces commercial risk. One possible hope for amelioration is the growing pressure on companies to increase the completeness and transparency of clinical trial data.<sup>45</sup>

**3.5. Challenges Associated with P.** The use of standardized names and accession numbers to enable the resolution of P is more common than for C entities in papers. Nevertheless, ambiguity can still be a problem. In GtoPdb, we cross-refer commonly used protein or gene identifiers, including our own NC-IUPHAR nomenclature. For many reasons, we use the UniProtKB accession number as an unequivocal, species-specific, primary identifier, which, for human proteins, is the Swiss-Prot curated entry. We also cross-refer the approved HGNC gene symbol as part of our HGNC/NC-IUPHAR collaboration. For those more familiar with NCBI annotation, a RefSeq NM or NP can also be used, but a gene ID is preferable. Where the paper specifies a protein complex, we need to resolve this to the constituent subunit IDs. Notwithstanding our efforts to resolve sequence identifiers, we are aware that *in vitro* activity studies are often done with forms of the protein target, where either the primary sequences are not identical to the Swiss-Prot canonical entry or various kinds of post-translational modifications have been introduced that could plausibly affect activity. For example, the protein could have a mutation used in a resistance screen, have a His tag left over from purification, Methionines may have been substituted to reduce oxidation sensitivity, the N-glycosylation could be variable according to the expression system, or, as in the cases of some proteases, both the signal and propeptide may be omitted from the expression construct. Here again, if notable deviations from wild-type are specified by authors, these are mentioned in curators’ notes and the reference is linked to the entry for further details.

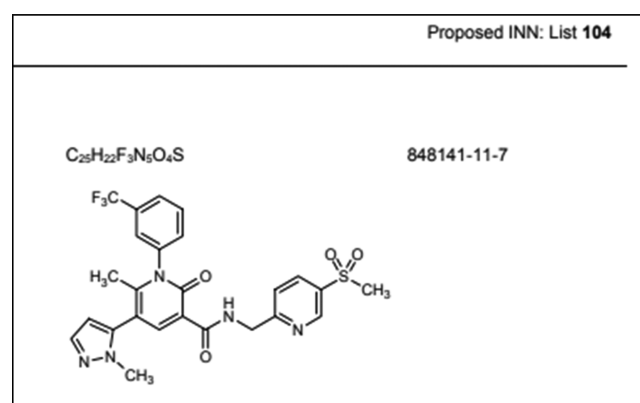
#### 4. ILLUSTRATION OF CURRENT c2p CHALLENGES

We can translate the general c2p challenges above into a specific example different to the BIA 10-2474. In this case, we selected “AZD9668: pharmacological characterization of a novel oral inhibitor of neutrophil elastase” as the primary citation D<sup>46</sup> for the lead compound alvelestat (LID 6476). Our c2p authenticity judgment was high for the following reasons: (a) the journal is the fifth-ranking for our citations, (b) the authors are from an established medicinal chemistry team in a major pharmaceutical company, (c) it is open access (OA), (d) as an older paper, the pharmacology has been corroboratively reproduced in other articles as well as AstraZeneca published patents, and (e) the

level of detail in the paper allowed unequivocal D-A-R-C-P assignments. For example, A-R was covered by the quote “AZD9668 had a high binding affinity for human NE ( $K_d = 9.5$  nM) and potently inhibited NE activity” (Table 1 from PMID 21791628). The calculated pIC<sub>50</sub> (IC<sub>50</sub>) and  $K_i$  values for AZD9668 for human NE were 7.9 (12 nM) and 9.4 nM, respectively”. In addition, Table 2 (from PMID 21791628) includes a log-transformed standard error of the mean as a pIC<sub>50</sub> of  $7.9 \pm 0.12$ . For P, “neutrophil elastase” can be resolved to UniProtKB accession P08246 and can also be identified as HGNC symbol ELANE, NCBI gene ID 1991, and GtoPdb TID 2358. However, we noted, as for other enzyme families with a history of alternative names, ELANE can potentially be confused with five members of the chymotrypsin-like elastase family (CELA1, CELA2A, CELA2B, CELA3A, and CELA3B). This includes what used to be called pancreatic elastase that now splits into the last three gene names in the list.

The assignment of n2s in this case had caveats that illustrate an important c2p issue.

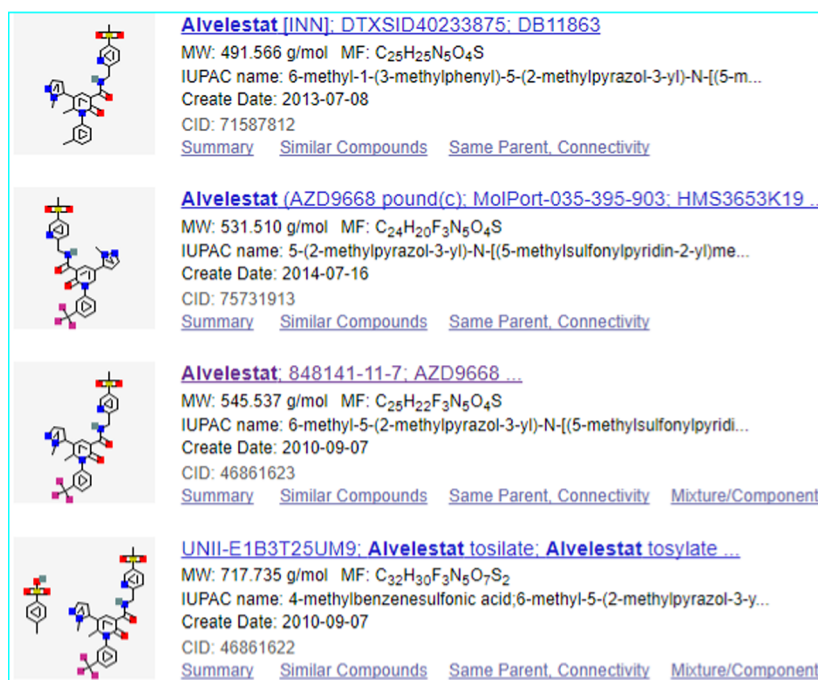
The information in the paper was consistent with the proposed INN: list 104 WHO drug information, Vol. 24, No. 4, 2010, where alvelestat was first declared as n2s, as shown in Figure 7.



**Figure 7.** Image for alvelestat from the INN proposal document. This included the IUPAC name: *N*-{[5-(methanesulfonyl)pyridin-2-yl]-methyl}-6-methyl-5-(1-methyl-1*H*-pyrazol-5-yl)-2-oxo-1-[3-(trifluoromethyl)phenyl]-1,2-dihydropyridine-3-carboxamide.

Importantly, OPSIN converted the IUPAC name into SMILES and InChIs that in turn allowed an unequivocal s2s match with CID 46861623. During curation, we also noted that this CID has clear submitter “majority vote” at 45 SIDs, thus supporting the parsimonious assumption that this n2s (i.e., for AZD9668 and the alvelestat INN) is correct. As a cross-check, using the highest stringency query of “alvelestat”[Completesynonym] in PubChem compound search (as a complete string match) returns only CID 46861623. However, a simple name search (that includes partial string matches) returned a surprising result (Figure 8).

The INN name query returns four distinct CIDs (each differing by molecular weight and InChIKey), which, from the PubChem chemistry rules, must have at least one submitter each with different n2s. Inspection shows that seven submitters for CID 172650821 chose the tosylate salt with the correct parent structure (albeit with two spellings of the counterion). It is unclear where this mixture originated, since this is not an official USAN salt name. However, we can establish that ChemIDplus was the first to assign it in 2013 against a patent extraction by



**Figure 8.** CIDs retrieved from PubChem with the term “alvelestat”. The SID counts for each of these in descending order are 3, 5, 45, and 7.

Thomson from 2010. Notwithstanding, it is neither clear how the three SIDs for CID 71587812 (ChemIDplus, DrugBank, and EPA DSSTox) omitted the trifluoride nor how the five submitters for CID 75731913 (including four vendors) dropped a methyl group. Statistics on the frequency of this type of n2s error in open sources (i.e., not just stereochemistry ambiguities) are difficult to come by. However, surveys of nonsystematic identifiers (such as INNs) matching more than one structure indicated a within-databases median of 2.5%, levels that, while not catastrophic, are still confounding for c2p.<sup>47</sup>

**4.1. D-A-R-C-P Limitations.** Although we find D-A-R-C-P mapping a useful summary of our curation process, we are aware of its limitations. The most important of these is the incomplete capacity (in terms of edge cases) to capture the full range of mechanistic nuances of c2p. We can exemplify some of these limitations as follows:

- (1) There are many examples of pharmacologically significant compounds with system perturbation read-outs but where the mmoa has not yet been elucidated (i.e., C may have a phenotypic A without a P). While lithium remains the classic case (LID 5212), we can also use the example of CCG-1423 (LID 6761). This inhibits RhoA transcriptional signaling via an unknown target but is a useful tool compound for disrupting this pathway in cancer.
- (2) The conversion of prodrugs to drugs is a well-established medicinal chemistry strategy but presents a curatorial dilemma. In some cases, we can record inhibitory activity for what is named as the approved drug in regulatory documentation (i.e., C has an INN), but the metabolite is more active by orders of magnitude (i.e., a “C-to-C” conversion relationship). Our entry for ACE (TID 1613) has 32 ligands including 13 prodrug/drug (lil/lat) pairs. There are other cases that, while not designed as prodrugs per se, generate active metabolites. Here, we provide cross-pointers, as for the two hydroxyatorvastatins from our atorvastatin entry (LID 2949).
- (3) Some important medicines are molecularly undefined substances and thus difficult to consistently map to database identifiers (i.e., “C” is not pure). In the classic case of heparin, we use a work around in the form of a defined structure (LID 4214) to represent the use of heterogeneous purified heparins in vivo. As a compromise, this allows us to add a useful “C-A-P” mapping.
- (4) Indirect mechanisms (i.e., where P is not the primary efficacy target). Heparin is also a good example in this case, where “C-P” is antithrombin III, but it indirectly acts as a thrombin inhibitor (although such cases are rare).
- (5) Imaging reagents. By definition, these are not therapeutic but are very important diagnostically. We can often find a reported “A-R-P” result but not always. Flortaucipir binding to tau is an example (LID 9100).
- (6) Single-target mapping has an obvious shortcoming where evidence supports the binding of C is to a constitutive heteromeric protein complex (i.e., consisting of one A-R to multiple P’s). While NC-IUPHAR has defined a set of complex targets for GtoPdb, there are exceptions where, based on reported data, we assign a single protein from the complex as the main binding partner. An example here is presenilin (PSEN1, TID 2402) against which nine  $\gamma$  secretase inhibitors are mapped. We made this choice to constrain our relationship matrix, which would otherwise be complicated with the additional four proteins in the  $\gamma$  secretase complex.
- (7) The other equally obvious shortcoming of single-target mapping is broadly termed polypharmacology. This (as one C against multiple A-R-P) can be divided into multiple efficacy-related targets and off-targets as liability-associated. Here again we veer toward stringency of only including quantitatively defined (multiple binding) targets and designating a data-supported primary efficacy target. The drug verubecestat (LID 8699) is an interesting example where we record that it is more potent against the

paralogue BACE2 than its primary target in Alzheimer's disease of BACE1.

- (8) Different physical forms of chemically identical active compounds. These are very important therapeutically (e.g., different crystallization polymorphs giving fast or slow release), but we do not index formulations.
- (9) New therapeutic modalities are on the horizon that will present data model challenges. We already have recently approved nucleic acid drug entries (i.e., where A and P are missing) as exemplified by Nusinersen (LID 9416). We also expect to see small molecules optimized for direct binding against defined nucleic RNA sequences where we will have to replace P with a defined RNA accession number.<sup>48</sup> Another area of development is large molecule–small-molecule covalent conjugates (hybrids), which are antibodies with cytotoxic payloads. Some are already approved, such as ado-trastuzumab (LID 6928), where, pending future options, we cross-point to the “warhead” as a separate entity. This type of problem has inspired global efforts directed toward the technicalities of database registration of complex biological reagents in general, including for regulatory requirements (e.g., the idea of macromolecular InChIs is being developed).

## 5. FUTURE OUTLOOK

**5.1. Overcoming Challenges.** After focusing primarily on challenges, we can move on to consider opportunities for c2p enhancements, in the sense of at least overcoming some of the challenges. However, we need to prelude this with an obvious question; has our task of D-A-R-C-P curation become any easier over the last 15 years? Stepping back to take our own perspective we would have to answer—not that much. Those easing effects that we have noted include the following: (1) We make a concerted effort to pass on curatorial problem-solving experience internally through changes of team personnel. This takes many forms but has an extent of formalization in our own FAQ, our team publications, extensive support from NC-IUPHAR committees, and the development of a sophisticated curation tool that encodes accumulated guidelines for filling in new database entries (i.e., we get better at what we do). (2) Since the completion of the human genome and the annotation efforts of resources such as HGNC and UniProt, protein target ambiguity from author descriptions has diminished (but by no means disappeared). Improved journal editorial guidelines have also helped. (3) The three tools mentioned above, OPSIN, Chemicalize, and OSRA, all appearing relatively recently, have proved very useful. (4) The expansion of PubChem (via the curated sources contributing to it, including our own) now provides a 95 million chemical structure substrate for cross-checking C from papers and patents.

**5.2. Emerging Opportunities for Connecting c2p.** The most significant of these would be large-scale, open disinterring of A-R-C-L-P relationships from pharmacology literature. However, there seems no immediate prospect of these five entities and attributes being surfaced and connected to the level of precision that GtoPdb is known for. We should also point out that, in general, selecting papers has not been rate limiting for GtoPdb expansion, but curator time certainly is. So, while GtoPdb will remain small and stringent as a strategic choice, it does raise the question of just how large the c2p corpus could be, with the crucial caveat of quality filtration if possible. As the figures above show, ChEMBL has the largest manual extraction

capacity of any public source. However, even with these resources, the combination of curatorial lead time and long release cycles gives a backlog of up to 2 years. With GtoPdb's smaller in-house curation team, new lead compounds and targets can be added within our 2 month release cycle.

Consideration of extraction scale leads naturally to the subject of text mining and its more advanced form as natural language processing. This has now developed to the point where it is certainly capable of recognizing and extracting D, A, R, C, and P from full text.<sup>49</sup> However, joining them up as D-A-R-C-P automatically is still challenging. In this context, it should be noted that professional biocurators can make such joins from a clearly laid out manuscript in a matter of minutes (but somewhat longer if the entities are confusingly arranged, including Markush-only structures, or results are partitioned into multiple supplementary data files). However, even as progress is being made in the precision of automated entity extraction from text, the position regarding publishers permissioning for text mining of the 4.3 million full text articles in PubMed Central and EPMC (recently expanded by ~2 million via the “unpaywall” initiative) remains unclear. This presents a paradox in that, since patents are free of licensing restrictions, SureChEMBL not only generates an open database of 19.2 million chemical structures automatically extracted from documents but also performs an extended bioentity mark-up within selected documents.<sup>50</sup>

So what future developments could drive the availability of D-A-R-C-P? We suggest, logically, a process needs to be widely adopted whereby authors specify both the entities and the key data relationships from their own papers. Since they carried out the scientific work, they are de facto the best stakeholder group to do this. There are three corollaries. First, iteration between authors and journals and biocurators would be necessary to formalize the output (including some level of ontology mapping), and second, collaboration with open databases is necessary to accept the input. Arguably, there is no fundamental technical impediment to something along these lines being introduced, so resistance to this seems more of a social problem.

So why has progress toward this been slow and what could accelerate it? There are many possible factors of which only a few salient ones can be considered here. The first to mention is that from a journal pilot study in 2008 for protein interactions, authors proved unenthusiastic about filling in even a relatively simple data sheet.<sup>51</sup> Thus, if even just “P-to-P” data seemed onerous to comply with, then full D-A-R-C-P mark-up would seem to present an even greater adoption barrier (especially where C could expand out to dozens of structures from SAR sets). Nonetheless, we have seen some progress in this direction, including our own experience of working with *Br. J. Pharmacol.* and *Br. J. Clin. Pharm.* to update guidelines for authors on reporting c2p. They are now required to provide GtoPdb identifiers and nomenclature for the targets and ligands which their article is about. More recently, we have introduced a system whereby novel key entities (i.e., not already in GtoPdb) are communicated to the team for consideration before publication. Consequently, those within our remit are added and the new identifiers provided.

Journals have a fundamental role to play in c2p by providing clear guidance to authors on the database identifiers and submission formats to use. This also ideally extends to supporting live-links to external database entries from within articles to help readers to find further information and context for the C and P's as described. However, the four journals mentioned above constitute a relatively small proportion of the

total coverage for GtoPdb and, beyond these, there are neither declarations nor even word on the grapevine suggesting that many more journals (e.g., from our Figure 4 list) are planning to formally introduce author mark-up of D-A-R-C-P. A comparative paradox here is that authors complying with sequence accession numbers as a condition of publication (as mentioned above) have now reached a submission rate of ~3500 per day. Nonetheless, there are simple steps authors can take, even when there is no specific journal mandate. As we have shown, providing public database identifiers for C and P is a basic requirement, and if C is novel, then providing structural information in a standard format recognized by open chemistry tools is crucial. Authors and journals will see the benefits of enhanced connectivity within the whole knowledge domain, increasing impact, and, ultimately, citations.

To conclude, while this perspective has focused on GtoPdb, it covers key points regarding the curatorial conversion of unstructured, or semistructured c2p data into structured database records. These can not only justifiably be termed “knowledge bases” but are also being merged into “big data” aggregations for data mining, including artificial intelligence approaches. The prospective advantages for the pharmacology, chemical biology, and drug discovery fields are thus huge.

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: cdsouthan@hotmail.com.

### ORCID

Christopher Southan: 0000-0001-9580-0446

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

C.S., E.F., and S.D.H. were supported by Wellcome Trust grants 099156/Z/12/Z from 2012-2015 and 108420/Z/15/Z from 2015-2018. J.L.S. and A.J.P. were supported by the British Pharmacological Society and the University of Edinburgh. The authors thank all past and present members of NC-IUPHAR and its subcommittees for their guidance and support. They also thank our curator alumni, in particular Helen Benson and Chido Mpamhanga.

## ABBREVIATIONS

GtoPdb, IUPHAR/BPS Guide to PHARMACOLOGY  
 c2p, chemistry-to-pharmacology  
 mmoa, molecular mechanism of action  
 D-A-R-C-P, document-assay-result-compound-protein target  
 D-A-R-C-L-P, document-assay-result-compound-location (within D)-protein target  
 i2s, image-to-structure  
 n2s, name-to-structure  
 s2s, structure-to-structure

## REFERENCES

(1) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617.

(2) Harding, S. D.; Sharman, J. L.; Faccenda, E.; Southan, C.; Pawson, A. J.; Ireland, S.; Gray, A. J. G.; Bruce, L.; Alexander, S. P. H.; Anderton, S.; et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: Updates and Expansion to Encompass the New Guide to

IMMUNOPHARMACOLOGY. *Nucleic Acids Res.* **2018**, *46*, D1091–D1106.

(3) Harmar, A. J.; Hills, R. A.; Rosser, E. M.; Jones, M.; Buneman, O. P.; Dunbar, D. R.; Greenhill, S. D.; Hale, V. A.; Sharman, J. L.; Bonner, T. I.; et al. IUPHAR-DB: The IUPHAR Database of G Protein-Coupled Receptors and Ion Channels. *Nucleic Acids Res.* **2009**, *37*, D680–D685.

(4) Southan, C.; Boppana, K.; Jagarlapudi, S. A. A.; Muresan, S. Analysis of in Vitro Bioactivity Data Extracted from Drug Discovery Literature and Patents: Ranking 1654 Human Protein Targets by Assayed Compounds and Molecular Scaffolds. *J. Cheminf.* **2011**, *3*, 14.

(5) Butler, D.; Callaway, E. Scientists in the Dark after French Clinical Trial Proves Fatal. *Nature* **2016**, *529*, 263–264.

(6) Yates, B.; Braschi, B.; Gray, K. A.; Seal, R. L.; Tweedie, S.; Bruford, E. A. Genenames.Org: The HGNC and VGNC Resources in 2017. *Nucleic Acids Res.* **2017**, *45*, D619–D625.

(7) NCBI Resource Coordinators; Agarwala, R.; Barrett, T.; Beck, J.; Benson, D. A.; Bollin, C.; Bolton, E.; Bourexis, D.; Brister, J. R.; Bryant, S. H.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2018**, *46*, D8–D13.

(8) Ireland, S. M.; Southan, C.; Dominguez-Monedero, A.; Harding, S. D.; Sharman, J. L.; Davies, J. A. SynPharm: A Guide to PHARMACOLOGY Database Tool for Designing Drug Control into Engineered Proteins. *ACS Omega* **2018**, 7993–8002.

(9) Schulz, J. B.; Cookson, M. R.; Hausmann, L. The Impact of Fraudulent and Irreproducible Data to the Translational Research Crisis - Solutions and Implementation. *J. Neurochem.* **2016**, *139*, 253–270.

(10) American Type Culture Collection Standards Development Organization Workgroup ASN-0002. Cell Line Misidentification: The Beginning of the End. *Nat. Rev. Cancer* **2010**, *10*, 441–448.

(11) Mullard, A. Reliability of “new Drug Target” Claims Called into Question. *Nat. Rev. Drug Discovery* **2011**, *10*, 643–644.

(12) Ekins, S.; Olechno, J.; Williams, A. J. Dispensing Processes Impact Apparent Biological Activity as Determined by Computational and Statistical Analyses. *PLoS One* **2013**, *8*, No. e62325.

(13) Baker, M. Reproducibility Crisis: Blame It on the Antibodies. *Nature* **2015**, *521*, 274–276.

(14) Curtis, M. J.; Bond, R. A.; Spina, D.; Ahluwalia, A.; Alexander, S. P. A.; Giembycz, M. A.; Gilchrist, A.; Hoyer, D.; Insel, P. A.; Izzo, A. A.; et al. Experimental Design and Analysis and Their Reporting: New Guidance for Publication in BJP. *Br. J. Pharmacol.* **2015**, *172*, 3461–3471.

(15) Corsello, S. M.; Bittker, J. A.; Liu, Z.; Gould, J.; McCarren, P.; Hirschman, J. E.; Johnston, S. E.; Vrcic, A.; Wong, B.; Khan, M.; et al. The Drug Repurposing Hub: A next-Generation Drug Library and Information Resource. *Nat. Med.* **2017**, *23*, 405–408.

(16) Perlman, R. L. Mouse Models of Human Disease: An Evolutionary Perspective. *Evol. Med. Public Health* **2016**, *2016*, No. eow014.

(17) Davenport, A. P.; Alexander, S. P. H.; Sharman, J. L.; Pawson, A. J.; Benson, H. E.; Monaghan, A. E.; Liew, W. C.; Mpamhanga, C. P.; Bonner, T. I.; Neubig, R. R.; et al. International Union of Basic and Clinical Pharmacology. LXXXVIII. G Protein-Coupled Receptor List: Recommendations for New Pairings with Cognate Ligands. *Pharmacol. Rev.* **2013**, *65*, 967–986.

(18) International Society for Biocuration. Biocuration: Distilling Data into Knowledge. *PLoS Biol.* **2018**, *16*, No. e2002846.

(19) Li, G.; Kamel, M.; Jin, Y.; Xu, M. K.; Mbuagbaw, L.; Samaan, Z.; Levine, M. A.; Thabane, L. Exploring the Characteristics, Global Distribution and Reasons for Retraction of Published Articles Involving Human Research Participants: A Literature Survey. *J. Multidiscip. Healthc.* **2018**, *11*, 39–47.

(20) Zhu, K.; et al. Sphingosylphosphorylcholine and Lysophosphatidylcholine Are Ligands for the G Protein-Coupled Receptor GPR4. *J. Biol. Chem.* **2005**, *276*, 41325–41335 (Retraction, 2005, 280, 43280).

(21) Taldone, T.; Zito, S. W.; Talele, T. T. Inhibition of Dipeptidyl Peptidase-IV (DPP-IV) by Atorvastatin. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 479–484.

- (22) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.
- (23) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Motow, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.
- (24) Southan, C.; Manchala, A. K.; Devidas, S. *Large-Scale Curation of Bioactive Chemistry from Patents and Papers: Excelra GOSTAR*; 2017, <https://www.slideshare.net/cdsouthan/largescale-curation-of-bioactive-chemistry-from-patents-and-papers> (accessed July 2018).
- (25) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045–D1053.
- (26) Clark, A. M.; Bunin, B. A.; Litterman, N. K.; Schürer, S. C.; Visser, U. Fast and Accurate Semantic Annotation of Bioassays Exploiting a Hybrid of Machine Learning and User Confirmation. *PeerJ* **2014**, *2*, No. e524.
- (27) Gardossi, L.; Poulsen, P. B.; Ballesteros, A.; Hult, K.; Švedas, V. K.; Vasić-Rački, Đ.; Carrea, G.; Magnusson, A.; Schmid, A.; Wohlgemuth, R.; et al. Guidelines for Reporting of Biocatalytic Reactions. *Trends Biotechnol.* **2010**, *28*, 171–180.
- (28) Orchard, S.; Al-Lazikani, B.; Bryant, S.; Clark, D.; Calder, E.; Dix, I.; Engkvist, O.; Forster, M.; Gaulton, A.; Gilson, M.; et al. Minimum Information about a Bioactive Entity (MIABE). *Nat. Rev. Drug Discovery* **2011**, *10*, 661–669.
- (29) Lane, J. R.; May, L. T.; Parton, R. G.; Sexton, P. M.; Christopoulos, A. A Kinetic View of GPCR Allostery and Biased Agonism. *Nat. Chem. Biol.* **2017**, *13*, 929–937.
- (30) Krippendorff, B.-F.; Neuhaus, R.; Lienau, P.; Reichel, A.; Huisinga, W. Mechanism-Based Inhibition: Deriving  $K_{\text{land}}$  and  $k_{\text{inact}}$  Directly from Time-Dependent  $IC_{50}$  Values. *J. Biomol. Screening* **2009**, *14*, 913–923.
- (31) Wang, Y.; Cheng, T.; Bryant, S. H. PubChem BioAssay: A Decade's Development toward Open High-Throughput Screening Data Sharing. *SLAS Discovery* **2017**, *22*, 655–666.
- (32) Southan, C.; Sharman, J. L.; Benson, H. E.; Faccenda, E.; Pawson, A. J.; Alexander, S. P. H.; Buneman, O. P.; Davenport, A. P.; McGrath, J. C.; Peters, J. A.; et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: Towards Curated Quantitative Interactions between 1300 Protein Targets and 6000 Ligands. *Nucleic Acids Res.* **2016**, *44*, D1054–D1068.
- (33) Southan, C. Caveat Usor: Assessing Differences between Major Chemistry Databases. *ChemMedChem* **2018**, *13*, 470–481.
- (34) Southan, C. Expanding Opportunities for Mining Bioactive Chemistry from Patents. *Drug Discovery Today Technol.* **2015**, *14*, 3–9.
- (35) Clark, A. M.; Williams, A. J.; Ekins, S. Machines First, Humans Second: On the Importance of Algorithmic Interpretation of Open Chemistry Data. *J. Cheminf.* **2015**, *7*, 9.
- (36) Hersey, A.; Chambers, J.; Bellis, L.; Bento, A. P.; Gaulton, A.; Overington, J. P. Chemical Databases: Curation or Integration by User-Defined Equivalence? *Drug Discovery Today Technol.* **2015**, *14*, 17–24.
- (37) Noor, M. A. F.; Zimmerman, K. J.; Teeter, K. C. Data Sharing: How Much Doesn't Get Submitted to GenBank? *PLoS Biol.* **2006**, *4*, No. e228.
- (38) Gilson, M. K.; Georg, G.; Wang, S. Digital Chemistry in the *Journal of Medicinal Chemistry*. *J. Med. Chem.* **2014**, *57*, 1137.
- (39) Scott, J. D.; Li, S. W.; Brunskill, A. P. J.; Chen, X.; Cox, K.; Cumming, J. N.; Forman, M.; Gilbert, E. J.; Hodgson, R. A.; Hyde, L. A.; et al. Discovery of the 3-Imino-1,2,4-Thiadiazinane 1,1-Dioxide Derivative Verubecestat (MK-8931)—A  $\beta$ -Site Amyloid Precursor Protein Cleaving Enzyme 1 Inhibitor for the Treatment of Alzheimer's Disease. *J. Med. Chem.* **2016**, *59*, 10435–10450.
- (40) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Model.* **2011**, *51*, 739–753.
- (41) Southan, C.; Stracz, A. Extracting and Connecting Chemical Structures from Text Sources Using Chemicalize.Org. *J. Cheminf.* **2013**, *5*, 20.
- (42) Filippov, I. V.; Nicklaus, M. C. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. *J. Chem. Inf. Model.* **2009**, *49*, 740–743.
- (43) Lloyd, I. *Pharma Projects Pharma R&D Annual Review*; Citeline, 2018.
- (44) Southan, C.; Williams, A. J.; Ekins, S. Challenges and Recommendations for Obtaining Chemical Structures of Industry-Provided Repurposing Candidates. *Drug Discovery Today* **2013**, *18*, 58–70.
- (45) Goldacre, B.; Lane, S.; Mahtani, K. R.; Heneghan, C.; Onakpoya, I.; Bushfield, I.; Smeeth, L. Pharmaceutical Companies' Policies on Access to Trial Data, Results, and Methods: Audit Study. *BMJ* **2017**, *358*, j3334.
- (46) Stevens, T.; Ekholm, K.; Granse, M.; Lindahl, M.; Kozma, V.; Jungar, C.; Ottosson, T.; Falk-Hakansson, H.; Churg, A.; Wright, J. L.; et al. AZD9668: Pharmacological Characterization of a Novel Oral Inhibitor of Neutrophil Elastase. *J. Pharmacol. Exp. Ther.* **2011**, *339*, 313–320.
- (47) Akhondi, S. A.; Muresan, S.; Williams, A. J.; Kors, J. A. Ambiguity of Non-Systematic Chemical Identifiers within and between Small-Molecule Databases. *J. Cheminf.* **2015**, *7*, 54.
- (48) Rizvi, N. F.; Howe, J. A.; Nahvi, A.; Klein, D. J.; Fischmann, T. O.; Kim, H.-Y.; McCoy, M. A.; Walker, S. S.; Hruza, A.; Richards, M. P.; et al. Discovery of Selective RNA-Binding Small Molecules by Affinity-Selection Mass Spectrometry. *ACS Chem. Biol.* **2018**, *13*, 820–831.
- (49) Krallinger, M.; Rabal, O.; Lourenço, A.; Oyarzabal, J.; Valencia, A. Information Retrieval and Text Mining Technologies for Chemistry. *Chem. Rev.* **2017**, *117*, 7673–7761.
- (50) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; et al. SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2016**, *44*, D1220–D1228.
- (51) Orchard, S.; Hermjakob, H. Shared Resources, Shared Costs—Leveraging Biocuration Resources. *Database* **2015**, *2015*, No. bav009.