29 March

**An open-access tool for designing drug control into engineered proteins.**


Centre for Discovery Brain Sciences, Deanery for Biomedical Sciences, University of Edinburgh, Edinburgh EH8 9XD, UK.

# Current address: Biomolecular Structure & Modelling Unit, Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London WC1E 6BT, UK

* Address for correspondence: Dr Christopher Southan, Centre for Discovery Brain Sciences, Deanery for Biomedical Sciences, University of Edinburgh, Edinburgh EH8 9XD, UK. cdsouthan@hotmail.com

**ORCID IDs:**

Ireland S.M., 0000-0001-8248-7614  
Southan C., 0000-0001-9580-0446  
Dominguez-Monedero A. 0000-0002-0146-0602  
Harding S.D., 0000-0002-9262-8318  
Sharman J.L., 0000-0002-5275-6446  
Davies J.A., 0000-0001-6660-4032

**Abstract**

A major challenge in synthetic biology, particularly for mammalian systems, is inclusion of adequate external control for the synthetic system activities. Control at the transcriptional level can be achieved by adaptation of bacterial repressor-operator systems (e.g. TetR), but altering the activity of a protein by controlling transcription is indirect and for longer half-life mRNAs, decreasing activity this way can be inconveniently slow. Where possible, direct modulation of protein activity by soluble ligands has many advantages, including rapid action. Decades of drug discovery and pharmacological research have uncovered detailed information on the interactions between large numbers of small molecules and their primary...
protein targets (as well as off-target secondary interactions), many of which have been well studied in mammals, including humans. In principle, this accumulated knowledge would be a powerful resource for synthetic biology. Here we present SynPharm, a tool that draws together information from the pharmacological database GtoPdb and the structural database, PDB, to help synthetic biologists identify ligand-binding domains of natural proteins. Consequently, as sequence cassettes, these may be suitable for building into engineered proteins to confer small molecule modulation on them. The tool has ancillary utilities which include; assessing contact changes between different ligands in the same protein, predicting possible effects of genetic variants on binding residues and insights into ligand cross-reactivity between species.

**Keywords:**
Protein engineering, pharmacology, druggability, protein binding, binding site, prediction.

**Introduction**

Synthetic biology is a technology for engineering of new biological functions through the construction of novel genetic networks to realize novel metabolic, signalling and developmental pathways\(^1\)-\(^5\). Some synthetic biological systems use only natural proteins (i.e. as represented by the Swiss-Prot canonical sequence for that species), and achieve novel function by combining proteins not normally found in the same cell or even the same species. Other systems involve the use of novel proteins, themselves typically including domains chosen from various natural proteins and coded into an engineered gene: an example is the SynNotch synthetic cell-cell signalling system\(^6\). In many applications, there is a clear need for synthetic biological devices to be subject to external controls, for example to create adequate safeguards, to exert temporal and/or spatial control on a particular system. This need is particularly acute when the device is intended to be used in the general environment or in a medical implant. At the very least, there needs to be a reliable means to shut the system down quickly, and much thought is being given to this problem\(^7,8\).

Most control systems used to date have operated by using small molecules to control gene transcription. Typically, they use antibiotic-sensitive transcriptional repressor proteins from
bacterial systems, the operator sites of which are fused to the promoter of the synthetic gene: the well-known tetR system is a much-used example. These systems work well but their effect on protein activity is very indirect, blocking transcription of further mRNA for a protein but not affecting existing molecules of the protein itself nor of the mRNA from which new protein molecules will be translated. Constitutive differences in mRNA half-lives can, however, limit this approach for particular proteins. Direct control of protein activity would be faster, which is why this dominates natural inter- and intracellular signalling. For synthetic circuits, control by rapidly diffusing small molecules would be particularly useful and several novel controls of this type have been constructed, generally by a laborious process of selection from large libraries of protein variants. As modulators of activity, small molecules have many advantages over alternative forms of experimental functional modulation, such as CRISPR, RNAi or antibody blocking. Principal of these are - (a) rapid action, (b) dose response can be used to vary the effects quantitatively, (c) reversal by wash-out d) use of equal and low-potency analogues with different chemotypes as specificity and reproducibility controls, (e) although less common, activators or agonists may be suitable for positive modulation (i.e. gain-of-function interventions), (f) allosteric modulators offer a different type of kinetic control, and (g) small molecules can be accurately measured both pre and post experiment (e.g. to monitor input dosing and metabolic degradation).

The need for pharmacological researchers to access data for the interactions between drug-like molecules and their protein targets has resulted in the production of a range of databases aligned to this general task, starting with BindingDB in 2001. Updates on these resources have recently been reviewed. These databases present valuable sources of information that might help synthetic biologists identify drug-protein pairs in which the drug binding site of the protein is small, and self-contained enough to be used as a 'module' that will confer drug control on engineered proteins. This would allow rapid and direct modulation of the activity of the protein without the lag times involved in transcription, translation and degradation. The use of an approved drug as the controlling ligand would bring the additional advantage in that safety aspects of clinical drugs, and their possible off-target side-effects, are generally well-established. This makes the approach particularly valuable if the synthetic biological constructs are eventually to be translated into in vivo, clinical or animal-agricultural contexts. However, attempting such module selection from large-scale chemogenomic databases such as BindingDB, ChEMBL or even directly from PDB would be challenging. Various
types of PDB abstractions such as the sc-PDB ligand binding database\textsuperscript{18} and PDBbind\textsuperscript{19} are also useful resources but have long update cycles.

To make navigating these complex datasets easier, we have created an open-access, web-based tool that integrates pharmacological and protein-binding information to present data on the drug-binding domains of selected proteins in a manner useful to synthetic biologists. The interface we designed supports a variety of searching and browsing strategies and facilitates the choosing of the most appropriate protein domain to be used as a controllable module for a particular purpose. This resource, SynPharm, has been based on GtoPdb, an expert-curated, open-access database by the International Union of Basic and Clinical Pharmacology\textsuperscript{20} for the following reasons:

1. It is embedded in an environment with an active experimental Synthetic Biology team. This means that the initial bioinformatic / \textit{in vitro} testing cycles are already in progress (and the latter will feed back to enhancements of the former).

2. GtoPdb has a rapid release cycle of quarterly or less and it is intended to synchronise SynPharm updates at the same frequency.

3. Relative to the larger resources our less broad-ranging but pharmacologically selective PDB mappings present much smaller sets for users to easily navigate but still capture approved drugs and clinical candidates.

4. Every ligand in SynPharm is expert-curated and activity-mapped even though this activity is not always explicitly referenced in the publication associated with the PDB entry.

5. This means our selected ligands are also manually identified as authentically binding to specific protein pockets rather than inorganic ions and/or heteroatoms from crystallisation reagents.

6. Partially due to SynPharm but also because of the increasing interest in new receptor and enzyme ligand structures in general, we have been recently enhancing our capture by triaging new human PDB depositions.

7. Beyond direct application to synthetic biology \textit{per se}, SynPharm has ancillary utility for GtoPdb users to explore ligand structures.

The Results section below presents the web pages that we have instantiated for SynPharm, the technical construction of which is described in ‘Methods’.
Results

Our ligand identification process identified 804 ligand-target interactions that were associated with at least one PDB code. Manually checking these interaction-PDB maps and rejecting duplicates gave a preliminary list of 768 interactions with associated PDB files. Among the interactions with structural data, 744 of the 768 (97%) interactions concern human data, with 15 (2%) rat, 8 (1%) mouse targets, and 1 *Plasmodium falciparum* target.

Our results established (not unexpectedly) that the distribution of interactions for which there is identifiable structural data is unevenly distributed between target families, as shown below in Table 1.

<table>
<thead>
<tr>
<th>GtoPdb target type</th>
<th>Targets with interactions (with or without structures)</th>
<th>Number of interactions (with or without structures)</th>
<th>Targets with structural data</th>
<th>Interactions with structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPCR</td>
<td>277 (16%)</td>
<td>9078 (52%)</td>
<td>29 (12%)</td>
<td>67 (11%)</td>
</tr>
<tr>
<td>Enzyme</td>
<td>755 (44%)</td>
<td>3518 (20%)</td>
<td>157 (64%)</td>
<td>365 (60%)</td>
</tr>
<tr>
<td>VGIC</td>
<td>127 (7.5%)</td>
<td>1408 (8%)</td>
<td>2 (0.8%)</td>
<td>3 (0.5%)</td>
</tr>
<tr>
<td>LGIC</td>
<td>66 (3.9%)</td>
<td>1027 (5.9%)</td>
<td>4 (1.6%)</td>
<td>4 (0.66%)</td>
</tr>
<tr>
<td>Other ion channel</td>
<td>47 (2.8%)</td>
<td>201 (1.2%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Catalytic Receptor</td>
<td>178 (10%)</td>
<td>992 (5.7%)</td>
<td>13 (5.3%)</td>
<td>40 (6.6%)</td>
</tr>
<tr>
<td>NHR</td>
<td>35 (2.1%)</td>
<td>523 (3.0%)</td>
<td>25 (10%)</td>
<td>104 (17%)</td>
</tr>
<tr>
<td>Transporter</td>
<td>120 (7%)</td>
<td>433 (2.5%)</td>
<td>1 (0.4%)</td>
<td>4 (0.66%)</td>
</tr>
<tr>
<td>Other Protein</td>
<td>99 (5.8%)</td>
<td>231 (1.3%)</td>
<td>16 (6.5%)</td>
<td>23 (3.8%)</td>
</tr>
</tbody>
</table>
Table 1: Representations of different classes of targets in GtoPdb (release 2018.1) that have any interaction data and that have useful structural data. Key to abbreviations: GPCR = G protein-coupled receptors; VGIC = voltage-gated ion channels; LGIC = ligand-gated ion channels; NHR = nuclear hormone receptors.

As is well known some target classes are more tractable to X-ray determination and consequently proportionally more highly represented with structural data. For example, Nuclear Hormone Receptor (NHR) interactions are particularly structure-dense in comprising 17% of the annotated sequences but just 3% of GtoPdb interactions overall. Enzymes are also over-represented in that just 20% of GtoPdb interactions involve enzymes, but 60% of those proteins with structural data. By contrast, Voltage Gated Ion Channels (VGICs) have just 3 annotated sequences (0.5%), compared with 1,408 (8%) total interactions. This bias reflects the inherent difficulties with structural studies of membrane proteins, although recent advances have lead an increase in the number of GPCR structures in the last few years, many of which include ligands\textsuperscript{21}.

Our process of compiling the SynPharm resource, detailed in Methods, is outlined in Figure 1. The output has been used to populate the home page designed to allow users to search the dataset by ligand or target protein (Figure 2).
Figure 1: The strategy used to produce a database of potentially useful interactions from known binding data.
Figure 2. The SynPharm home page at http://synpharm.guidetopharmacology.org/

Users may browse the site without a specific molecule in mind or alternatively take as their starting point identifying any ligand-binding segment of a protein that might be transferrable to an engineered protein in their project. In this case, clicking on the “Sequences” link without entering a search term lists all target proteins in the list of potentially useful pairs described above. This list can be browsed as shown in Figure 3.
In Figure 3 targets have been ordered by length of ligand-binding segment but they can be ordered by any of the columns by clicking on the table headers. These metrics can provide useful first-pass information to prioritise more detailed analyses. Selecting any target brings the user to its sequence page. At the head of this page is a three-dimensional visualisation of the target chain bound in complex with the ligand, with the binding segment itself highlighted in green to show its context within the original chain (Figure 4 A).
Figure 4: Examples of the types of structural display found on the sequence details page for human beta-secretase 1 in complex with the ligand AMG-8718 (sequence ID 84541). Panel A shows the 3D structure interactive viewer where the binding segment is highlighted in green, the rest of the target protein in white, and the ligand is shown in stick view. Panel B shows the residue distance matrix. The distance between any two residues in the target chain is denoted by colour – green to red – and, on desktop screens, hovering over any pixel will provide an exact numerical distance in Angstroms of the relevant residues. White portions denote residues missing from the PDB file of origin. The dotted line indicates the binding sequence. Panel C shows the binding portion of the sequence. The arrows allow extending
outwards along the rest of the sequence. Finally, panel D shows a zoomed-in section of the feature viewer. Binding residues are shown in context with secondary structure elements (α-helices and β-strands) and hydrophobicity over the peptide sequence.

The views in Figure 4 provide a rapid visual indication of how independent of the other features of the protein the segment's structure is likely to be: the more independent the structure, the more transferrable to other proteins the segment is likely to be. Visualization of the structure uses the JavaScript PV protein viewer\textsuperscript{22}. In addition to showing the structures, the sequence pages present metrics such as proportional chain length and contact ratio (used as a rough measure of likelihood that the sequence will fold correctly by itself, as it is a measure of ‘domain-likeness’: higher is more promising). The GtoPdb affinity data for the specific ligand-target interaction are also provided. Each sequence also has a residue distance matrix (Fig 4B), which depicts the distances between any two given residues in the binding chain, with the bind sequence itself highlighted with a black dotted line. This is to give a sense of the globularity of the sequence within the chain, and how compact it is.

There is also a feature viewer (Fig 4D) for each sequence, which utilises the biojs-vis-protein features viewer\textsuperscript{23}. In addition to binding residues and secondary structure elements, the feature viewer also maps hydrophobicity along the bind sequence, using the Kyte-Doolittle measures of amino acid hydrophobicity\textsuperscript{24}. There are extensive search functions for identifying sequences or ligands by various metrics. All ligands have links back to GtoPdb and a subset of their data are available directly on the SynPharm page, particularly molecular data and clinical approval information. These were chosen because they may be relevant to a researcher when picking a molecular switch inducer, but the full range of pharmacological data is accessible via the link back to GtoPdb. This can be illustrated for BACE1, an aspartyl protease drug target for Alzheimer’s disease\textsuperscript{25}. In Figure 5 a section of the BACE1 entry is shown and Figure 6 shows a sequence alignment of the extracted ligand interaction sections.
Figure 5. A snapshot from the GtoPdb BACE1 target entry http://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=2330. The eight SynPharm entries are indicated in the upper panel. From the total ligand entries in the lower panel, inhibition values are displayed for two of the SynPharm ligands, AZ-4217 and AMG-8718.
**Figure 6:** Differences in the contact residues extracted for the eight BACE1 ligands. The eight SynPharm sequences (in descending order) are 84891, 78993, 78985, 78477, 78900, 78917, 78921, 78923.
82636, 78987 and 84541. The latter (lowermost in Figure 6) is for AMG-8718 as shown in Figure 3. As for the SynPharm display the upper case letters indicate ligand contact residues.

A number of SynPharm advantages can be discerned from the BACE1 example, particularly since it is one of the most intensively perused drug targets (reflecting the massive unmet need for effective Alzheimer’s treatments). Metrics in support of this are that no less than 364 BACE1 structures are in PDB (nearly all with ligands), 11 of which were deposited in 2017. The ChEMBL 23 human BACE1 entry is linked to 6846 structures with some level of activity-mapping. The GtoPdb BACE1 entry maps 21 quantitative ligand interactions with a focus on clinical candidates and stringently selected research leads. Of these, 10 are in PDB and five are not in ChEMBL. From the 10, the eight indicated in Figs 4 and 5 have passed the SynParm extraction triage (described in Methods) and, as multiple ligands for the same target, provide a useful calibration. For example, the alignments shown in Fig. 5 indicate explicit differences between ligand interaction residues for the set even though alignment of the sequence sections indicates they are binding to the same pocket. We can note that all eight interact with Tyr 132 whereas only 82636 interacts with Glu 134 and Gly 135 while Tyr 137 only interacts with 78987 and 84541. These differences may be spacially minor (i.e. possibly only just outside the 5 Å limit used by SynPharm) but can nonetheless be useful. Even more useful to the synthetic biologist is to compare the overall length of the contiguous binding section for particular ligands. In Figure 5, we can see that five sequences extent out to Ala 396 as the ultimate interaction point. However, the results also indicate that only extending to DSGTT (or just past it in practical terms) may be sufficient.

While SynPharm should be sufficient as a stand-alone tool there are clearly external resources that complement it. The most obvious of these are RCSB PDB and PDBe, both of which both of which have complementary features for visualising ligand structures in proteins. In addition, we would recommend PDBSum for the display of a 3D complex where the LIGPLOT displays offer a useful complement to the SynPharm sequence mark-up. Another utility (accessible from GtoPdb out links) is that, via ranking by IC$^{50}$ in ChEMBL, we can find a 0.2 nM inhibitor (not in PDB) that would be of interest as potent control compound to be tested in an in vitro context (CHEMBL3394218, PubChem CID 118725974: this will be curated into GtoPdb for a forthcoming release).
Discussion

The idea of adding regulation into engineered proteins has already proved itself useful in a variety of contexts. A famous example is the addition of the tamoxifen-sensitive ERT2 domain into Cre recombinase to create a drug-inducible gene excision, allowing experimenters to remove gene function from an experimental animal at a time of their choosing\(^\text{27}\). This has been used in a variety of applications. Some have used the system to genetically mark cells for lineage tracing in development\(^\text{28,29}\), disease\(^\text{30}\), and regeneration\(^\text{3, 31}\). Other applications have used ligand-inducible Cre to examine gene function by removing it from a cell only at a chosen stage of development\(^\text{32,33}\). Induced Cre-mediated recombination has also been used to create sarcomas in model animals for the purposes of studying tumour development\(^\text{34}\). The technique has even been used in anatomical studies, deliberately sub-optimal doses of inducer being used to mark only sparse neuronal cells, allowing their detailed morphology to be studied in otherwise unlabelled tissue\(^\text{35}\). The use of photo-caged oestrogen adds an opto-genetic dimension to a version of the system using Cre-ER instead of Cre-ERT2, allowing light to be used to activate Cre-ERT2 in specific cells\(^\text{36}\).

The addition of ligand control is not limited to Cre. A similar technique has been used to add the ER domain to Snail, to study the role of that mediator of epithelial-mesenchymal transitions in controlling fibrosis in adult kidney disease\(^\text{37}\). The impact of these examples provides ample evidence that engineering control into proteins can be useful. It is our hope that the tools described here will be useful in the construction of further examples, broadening the range of ligands that can be used for this type of work.

Several caveats should be taken into consideration with our approach. One of these is the necessary restriction to contiguous sections of sequence: However, it is well known that overall binding energies are likely to have at least some contribution from long-range secondary structure interactions within the entire protein structure. Thus, the binding sequence cassette not only needs to fold correctly within the engineered host sequence, but it may also have a lower binding constant and altered kinetics (e.g. $K_{\text{on}}$ and/or $K_{\text{off}}$) compared with the full length native counterpart. This also means that the discontinuous binding sites characteristic of receptors, ion channels and transporters are largely excluded from our data.
harvest (but the associated ligands are not necessary ruled out for synthetic applications). Another caveat is that the inherent bias of GtoPdb ligand choice towards the literature associated with pharmacological activity in a human clinical context where possible: This selection is not optimal for ligand binding cassettes. Notwithstanding, for in vitro synthetic biology applications, complementary data can be explored, including searching for very potent inhibitors that are neither in GtoPdb nor the PDB but have a high likelihood of binding the same sequence section (and this could be supported by structural superimposition and/or docking experiments). We note also the caveat that the nesting-in of active site sections, by definition, could endow the host protein with enzyme activity. In this cases it should be possible to abolish such unwanted properties by mutating active site resides that are not major contributors to the binding energy.

We can point out utilities of SynPharm that extend beyond practical applications to synthetic biology per se. Firstly, the entries simply act as a convenient flag to users for the existence of PDB structures. Secondly, there is increasing interest in the effects of protein sequence variants that affect protein function in pharmacologically significant ways, for example they may affect patient drug responses if substitutions are found in the SynPharm sequences. Thirdly, by adding rodent or other model organism sequences to the sequence alignments shown in Fig.5, possibly complemented by homology modelling, some insight can be gained into orthologous cross-reactivity of ligands that could be experimentally tested.

The work described in this paper has resulted in a new, publicly available web resource mainly designed to help synthetic biologists to engineer pharmacological regulation into their proteins. We would be pleased to hear from other teams who would like to use this and we may be able to assist in cross-checking complementary sources to expedite their choices.

**Methods**

We used a sequential bioinformatic strategy for identifying ligand-binding sequence sections potentially useful to synthetic biology (Fig 1). Stage 1 was a screen for targets in GtoPdb for which any structural ligand binding data were available in the form of RCSB PDB files. This screen was performed by using GtoPdb web services to request PDB codes for each ligand associated with a target in GtoPdb (2018.1 release). To obtain further structural data on this first set of potentially interesting interactions, the RCSB PDB web services were queried with
information from GtoPdb. For each ligand-target interaction, PDB codes associated with ligands were obtained by searching on ligand code, name, SMILES, InChI, or peptide sequence. PDB codes associated with targets were obtained by searching the RCSB PDB web services using UniProtKB accessions.

Stage 2 was to identify amino-acid residues on the target that mediate each of the ligand-target binding interactions. The residues that mediate the ligand binding were identified by either using the information in REMARK 800 and SITE records of the relevant PDB file or, if no such records exist, by selecting all residues with atoms within 5 Angstroms of a ligand atom (ignoring hydrogen atoms). The binding sequence was then defined as the segment of the protein chain that contained all the ligand binding residues, for example a segment between amino acids 30 and 45 of a protein chain. If binding residues were on more than one peptide chain of a multi-peptide target protein complex, the interaction was rejected as not being useful for the purposes of protein engineering. Interactions were also rejected if more than 5% of the residues in the chain are ‘missing’ – that is, not observed in the PDB file (according to REMARK 465 records). This was the most frequent reason for discarding candidates. Stage 2 cut the list of potentially useful interactions down to 610 sequences. This is a relatively small proportion (3.5%) of the number of interactions in the 2018.1 release of GtoPdb, a reflection of the small number of target-ligand interactions for which there exists structural data.

Stage 3 associated certain metrics with each sequence. These were; (i) its length as a proportion of the original chain, (ii) its ‘contact ratio’ – defined as the ratio of internal contacts (all non-hydrogen atom pairs within the sequence within 5 Angstroms of each other, excluding atoms within two covalent bonds of each other) to external contacts (all non-hydrogen atom pairs between the sequence and the rest of the chain, less than 5 Angstroms). In cases where an interaction had multiple PDB maps and so multiple potential sequences to represent it, we selected those with the smallest length proportional to their original chain length as the most likely to be useful for engineering purposes. The system also allows manual selection of an interaction-PDB map if this is required.
The functions for accessing the GtoPdb web services have been bundled into a standalone Python library called pyGtoP, and the code for parsing PDB files and identifying the various elements within them (used in sequence construction) has been bundled into a Python PDB parser called molecuPy. Both are open source projects viewable on GitHub. The scripts which used these new libraries to do the work described above, as well as the code for the database and web interface itself, are also open source and viewable on GitHub in the synpharm repository.

Construction of a web interface.

Our aim was to make the data available in a useful format to synthetic biologists, in the form of an easy-to-use web page. We have therefore stored the data in a PostgreSQL database, with a separate staging database to make future updates easier. This is connected to a web page using a Java (Oracle Corporation, Redwood City, CA, USA) web application installed on an Apache Tomcat web server (The Apache Software Foundation), and the web page is open access at.

Acknowledgments. SMI and ADM were supported by BBSRC grant BB/M018040/1, CDS and SDH by Wellcome Trust grant 108420/Z/15/Z and JLS by the British Pharmacological Society.
References


22 https://github.com/biasmv/pv

23 https://github.com/onursumer/biojs-vis-proteinFeaturesViewer


38 https://pygtop.readthedocs.io/en/latest

39 github.com/samirelanduk/pygtop, github.com/samirelanduk/atomium/tree/molecupy1.0.5

40 github.com/samirelanduk/synpharm

41 https://www.postgresql.org/

42 synpharm.guidetopharmacology.org