# The Functional Therapeutic Chemical Classification System

# Abstract

A drug can usually play more than one role in the human body. This characteristic known as poly-pharmacology results from the complex interaction of the therapeutic chemical with multiple molecular targets, themselves carrying multiple biological functions. Despite being unwanted, the side-effects resulting from poly-pharmacology can sometimes reveal compounds with unexpected beneficial actions; in such cases, the drug can be repurposed, namely re-indicated for a new disease category. Several computational approaches aim at identifying potential molecules for repurposing using various descriptors, such as the chemical structure or gene expression patterns for instance.

A step halfway between the therapeutic indication and the molecular structure is the mode of action (e.g. *pro-apoptosis agent*). This concept has not yet been taken into consideration for drug repurposing studies, yet we believe that it can help to characterise the multiple roles a drug carries. We used description logics and their implementation as the Web Ontology Language (OWL) in order to formally define the meaning of over 20’000 terms describing various modes and mechanisms of action. Based on an integration of data coming from DrugBank, Uniprot, the Gene Ontology and related protein annotations (GOA), we have automatically assigned over a thousand of approved drugs into the mode of action categories.

The resulting new resource is called the Functional Therapeutic Chemical Classification System (FTC) and was further evaluated against the content of the traditional Anatomical Therapeutic Chemical Classification System (ATC). Compounds in the FTC belong to multiple mode of action categories, reflecting the various potential roles they can play. We illustrate how the resource can be used to generate drug repurposing hypotheses, using Alzheimer’s disease as a use-case. A web application built on the top of the resource is freely available at <https://www.ebi.ac.uk/chembl/ftc>.

# Author Summary

Developing a new drug to treat a particular disease is a long process which takes at least a dozen of years. Recently, numbers are showing that less and less drugs are reaching the market, most of them failing during expensive clinical trials, where their toxicity and efficacy is tested on real patients. In order to overcome this bottleneck, a strategy called drug repurposing consists at finding new uses for already approved and therefore safer compounds. Encouraged by successful experimental results from the past, computational methods have been recently developed to look in a systematic fashion at the drug repurposing issue. We present with this work a new approach consisting at defining in a mathematical way the concept of *mode of action.* Briefly, the mode of action of a drug characterise the role a molecule carries in a living body (e.g, *anti-blood coagulation agent*). The resource we present here is called the Functional Therapeutic Chemical Classification System and consist of large number (20’000) of new mode of action terms. We have then integrated the data from various databases to automatically assign approved compounds to the various mode of actions terms. The resource we built to understand better the multiple roles a drug can have and to facilitates the generation of drug repurposing hypotheses, which can be further experimentally tested.

# Introduction

According to Wikipedia (cite - date), drug repurposing is the *application of known drugs and compounds to new indications (i.e., new diseases)*. When administered in a living body, a pharmaceutical agent can indeed play various roles and affect different biological processes. Accurately identifying these different functions helps to predict the potential side-effects a drug could have and can also lead to interesting repurposing opportunities. For instance the sildenafil was initially developed against angina and has been repurposed towards erectile dysfunction during the clinical trials (cite); an unforeseen activity of the molecule was provoking prolonged erections in male patients and the drug has been re-indicated accordingly. Approved compounds are attractive because they have by definition already successfully passed clinical trials, where most of the drugs are failing. Moreover, such molecules have been extensively studied and lots of knowledge exists about them.

The current trend in the field is at predicting repositioning instances using computational methods (cite). The traditional approach compares compounds against a physical descriptor such as the molecular structure (cite). Other methods characterise the drugs on more abstract levels, such as the gene expression signature (cite) or via the reported side-effects for example (cite). These approaches all have in common to look at similarities against one or more descriptors, similar compounds being inferred as repurposing hypotheses.

Another feature of particular interest to describe drugs is the *mode of action (MoA)*. *The MoA describes a functional or anatomical change, at the cellular level, resulting from the exposure of a living organism to a substance (cite wikipedia)*. For instance words such as *transcriptional regulation agent* or *anticoagulant* are defining MoAs and characterising the roles of a certain type of biologically active chemicals. The MoA is the central concept linking a chemical structure to a set of biological effects. Intuitively, the indication of a drug logically depends on its MoA.

Despite being a widely spread notion in drug discovery, the MoA has not been used yet as descriptor for repurposing analysis. First, in order to be exploited computationally it requires to be clearly defined. Indeed, a MoA is an abstract concept and different scientists can have various interpretation of its meaning. MoAs are words or categories, it is not possible to represent them straightforwardly with values and numbers like one can do for a 3D molecular structure or for a gene expression profile. Nonetheless, the meaning of a concept can be formalised with controlled vocabularies and ontologies (cite). Originating from description logics, these frameworks help to define the mathematical meaning of symbols and strings of characters. In an ontology or knowledge base, concepts are organised and linked following the logical type of relation they have between them. In the biomedical domain, the most well-known example is probably the Gene Ontology (GO) (cite). Biological processes and molecular functions are manually curated in this resource and the meaning of terms is specified by the relations linking two GO terms.

MoAs definitions are present in other classifications such as the Medical Subject Headings (cite) or the Chemical Entities of Biological Interest (cite) for example. The Anatomical Therapeutic Chemical Classification System (ATC) also describes to some extend the action of drugs at the anatomical level. All these resources are valuable for the community as a source of carefully manually curated information. Moreover, the categories described in these classification systems are sometimes used to annotate drugs. For instance the compound sildenafil (CHEBI:9139 or MeSH:C101426) has a vasodilator agent role (CHEBI:35620 or MeSH:D27.505.954.411.918), this assertion being manually done by a curator.

The classifications mentioned previously are not specially designed for drug reprofiling, they report on purpose only the well-known and major MoAs of chemical compounds. Secondary poly-pharmacology is also not necessarily well covered, yet it would be the best way to predict repositioning opportunities. In our context, an ideal knowledge base would feature the known MoAs of a drug as well as some predicted one to be tried in experiments. In order to address this hypothesis, we have implemented the Functional Therapeutic Chemical Classification System (FTC), which will be presented in this manuscript. The FTC is automatically built by leveraging the content of various biomedical databases using description logics and automated reasoning. Over 20’000 new MoA categories are defined in the resource, further populated with approved drugs using the Web Ontology Language (OWL) in combination with a reasoner. Most of the drugs are present in multiple FTC categories, reflecting the various roles a compound can play inside a biological system. The resource was evaluated against the ATC, traditional classification scheme introduced before. We present as well some preliminary analyses over the data, by looking at the relation between the MoA and the indication of a compound using semantic similarity. Finally, we illustrate also how the FTC can be used as a pharmacological toolbox to analyse drug repurposing for Alzheimer’s disease.

# Results

The knowledge base behind the FTC is built by integrating information coming from various sources. The GO terms serve as template to create the FTC categories describing the MoAs; DrugBank provides the links between drugs and their protein targets. Uniprot maps targets to their respective GO annotations. Drugs are further assigned into mode of action categories based of the OWL constructs and axioms defined in the FTC (see material and methods for details). A reasoner, program capable of understanding such axioms, performed this task. It took a few seconds on a standard laptop to classify the knowledge base using the ELK reasoner (cite). Other OWL reasoners have been tried (e.g, Hermit (cite), Pellet (cite), etc...) but none of them were suitable for this purpose; the classification time was always superior to a few minutes (data not shown). The FTC has a taxonomic structure as illustrated on Figure 1, which arises when the reasoner classifies the knowledge base. Categories can have multiple parents and multiple children. The reader can browse and access the content of the FTC online at <https://www.ebi.ac.uk/chembl/ftc>.

Figure 1: Parent categories to the FTC class *Pro-fibrinolysis agent* (FTC\_P0042730). The classification is a direct acyclic graph where categories are describing increasingly specific concepts.

In total there are 1280 approved drugs perturbing 1264 human protein targets. The FTC introduces 23’353 new categories describing the mode and mechanism of action of therapeutic compounds. 4289 of these categories are related to biological processes and 19’064 to the molecular functions. A summary of the metrics behind the latest build is available online at <https://www.ebi.ac.uk/chembl/ftc/evaluation/>. Out of all FTC classes, 1432 categories (≈ 6%) directly contain at least one approved drug. This number increases up to 2532 (≈ 11%) when direct and indirect drugs are considered. FTC categories not containing drugs (e.g, FTC\_A0001771 - *Anti-immunological synapse formation agent*) represent modes and mechanisms of action for which no approved compounds exist already or that have not been identified as such in the FTC.

## Evaluation

The content of the FTC has been evaluated against the information present in the ATC. The motivation behind this step is to have an estimation how the automatically-built classification (FTC) compares to a manually curated one (ATC). In the context of this work, we have considered the ATC as a gold standard against which the FTC has been assessed. In practice, these two resources have different goals therefore the evaluation should be interpreted accordingly. The full methodology behind the evaluation is available in material and methods. Briefly, some classes present in the FTC have an equivalent meaning to some classes present in the ATC. An *evaluation point* was manually asserted all the time this situation happened (total of 68 evaluation points). Then the set of drugs present in both classes for each evaluation points are compared and some metrics can be derived from it. For example, the FTC category *Anti-hydrogen:potassium-exchanging ATPase activity agent* (FTC\_A0008900) has been manually asserted as equivalent to the ATC category *Proton pump inhibitors* (A02BC). A summary of this evaluation point is furthermore available online at <https://www.ebi.ac.uk/chembl/ftc/evaluation/FTC_A0008900>.

Out of the 1280 DrugBank compounds present in the FTC, 1134 are also present in the ATC, therefore only those are considered. The evaluation points are covering a total of 471 unique drugs. 277 compounds are true positives, meaning that given an evaluation point, they are present in both FTC and ATC categories. The evaluation point discussed in the previous paragraph (*proton pump inhibitors*) is an example where all the drugs (omeprazole, esomeprazole, pantoprazole, lansoprazole, rabeprazole) are present both in the FTC category and in the corresponding ATC class. The total numbers of compounds present in an ATC class but not in the corresponding FTC category (false negatives) is 36. Finally, 279 compounds are present in a FTC class but not in the corresponding ATC class, such agents are the false positives. From these values, it is possible to derive a recall of 88%; this percentage indicates that the automatic build of the FTC covers a good portion of the content already present in the ATC. The precision of 50% shows that the FTC contains for a given MoA much more drugs than the equivalent ATC categories. This characteristic is not necessarily a drawback, when looked at from a drug repurposing perspective: Some of the false positives can be interpreted as drugs that might be suitable for the therapeutic ATC category, yet not indexed as such.

The high recall value supports the idea behind the automated build of the FTC: The data from different repositories, funded and curated in parallel, can be integrated to automatically create a new resource. This new resource (FTC) containing most of the known information present in an external gold standard (ATC) and relying on description logics to leverage the native information. The design of the knowledge base generates some supplementary information (false positives), which could be considered as repurposing opportunities under the assumptions behind the axioms capturing the formal concept of MoA. Note that it is currently not possible to differentiate between genuine false positives and actual reprofiling hypotheses just based on the evaluation.

## Exploration

The FTC was designed to assist drug repurposing analyses by explicitly representing the poly-pharmacology of approved drugs. There are different ways to explore the content of the resource. As seen in the previous section, it is possible to derive reprofiling opportunities based on the evaluation. Another option is to look at a similarity metric derived from the tree structure of the FTC. The user can also start with a discrete MoA category and from this look at the list of potentially active approved drugs. In this section we exemplify these different types of analysis in order to demonstrate the potential of the resource.

### Poly-pharmacology spectrum

When administered in a living body, a chemical can produce a variety of phenotypic effects. The more is known about the molecular targets of a compound and their physiological roles, the more opportunities there are to re-orient the drug into doing something new or to prevent an adverse condition that could appear when administered. The therapeutic agents described in the FTC can have several MoAs, representing the poly-pharmacology born by the approved compounds. Figure 2 illustrates the poly-pharmacology spectrum by showing the distribution of number of MoAs per compounds. This number ranges from one to more than eighty for certain drugs, when indirect superclasses are considered. Not all the MoAs are biologically relevant, some FTC categories are particularly abstract (e.g, *Anti-biological process agent*) yet they represent discrete categories inside which the drug belongs with an explicit and clear meaning. These discrete MoAs are a good starting point to understand what a compound can do when administered in a human system. The human readable definitions associated to each FTC category help furthermore the user to choose the adequate MoA for the desired outcome.

Figure 2: Distribution of the number direct and indirect MoAs per drug (FTC categories inside which a compound has been classified). Means are indicated with a solid line. When considering indirect categories, the distribution range is wider.

### Drugs with similar mode of actions have similar indications

The list of MoAs attributed to a drug can be exploited as a descriptor for the therapeutic agent. A chemical can indeed be described based on physical parameters (structure, reactive group, etc...) but it can also be defined from more abstract notions, such as its role and function in the body. Because of the tree structure of the FTC it is possible to derive some similarity metrics over the MoAs, as performed over the GO and related gene product annotations (cite). The underlying heuristic is to assume that the closer two entities are in a taxonomy or ontology, the more similar they are. We used a straightforward approach derived from the Jaccard index (see method for more details) in order to compare approved drugs based on the similarity of their MoAs. For instance, the similarity between two compounds present in the same FTC category is 1 (maximum). The similarity between an *anti-blood coagulant* and *pro-blood coagulant* is 0.29, reflecting the fact that such compounds are dissimilar in regards to the outcome of their biological effect. As the MoA is intuitively expected to be the central concept leading to the indication of the drug, we expected that on average, drugs with similar MoAs would be indicated towards similar therapeutic areas. The heat map presented in Figure 3 shows a pairwise comparison of all the drugs of the FTC based on their relative MoA similarity. The compounds are further grouped by therapeutic indications as defined by the ATC. Only the first level of the ATC is considered on Figure 3 (see Supplementary Figure 1A for a two ATC level granularity).

Figure 3: Pairwise comparison of mode of action similarities. The similarity descriptor ranges from 0 (not similar - white) to 1 (identical - black). Drugs are grouped according to their first ATC level (colours on the side). For instance, the compound reteplase (DB00015) has the ATC code B01AD07, which appears as “B” (dark orange) on the plot. The average similarity of drugs present in the same therapeutic category is significantly higher on average when separately compared to all other indications.

The heat map displayed in Figure 3 reveals some square patches around the central diagonal. The overall similarity appears higher when compounds from the same ATC group are considered. A statistical analysis (cf material and methods) revealed that the average similarity of compounds belonging to the same ATC category is significantly higher than when compounds from different categories are compared. This result supports the idea that drugs with similar MoAs have similar indications. Note that the mean of the similarity values was considered for the analysis, some outliers are also present in the map, which can be interpreted as repurposing hypotheses. These outliers agents have indeed similar MoAs, yet they belong to totally different therapeutic areas and are used for different purposes according to the ATC. We are currently further analysing such cases in a systematic fashion. The reader can also look at Supplementary Figure S1A to observe a similar association behavior when two levels of the ATC are considered (no statistical analysis performed). Supplementary Figure S1B re-uses the same data as Figure 1 (one ATC level) but with a clustering function apply to it (hierarchical clustering - manhattan distance) in order to reveal functional clusters of compounds. Finally, Figure S1C shows the distribution when compounds are sorted based on their identifiers (≈ random); no visual pattern are identifiable in this case. Taken together, these results emphasize that the MoAs as defined in the FTC are indeed on average associated with the therapeutic indication of a drug. This statement supports the validity of representing the MoA in a formal, computer understandable way in order to address drug repurposing issues.

Supplementary Figure S1: Pairwise comparison of mode of action similarities. The compounds are grouped by ATC code (2 levels - legend not shown) on panel A. A hierarchical clustering step (based on manhattan distance) was applied on panel B. The dendrogram reflects the original structure of the FTC. Panel C shows the compounds sorted by DrugBank identifiers only (≈ random). The colours displayed on panel B and C correspond to one level ATC categories and are indexed in the legend panel.

### Drug repurposing hypotheses for Alzheimer’s disease

Previous sections have introduced the reader to the content of the FTC; the classification was first evaluated against the ATC, considered as a resource of reference. Then followed an exploration of the data, emphasizing the tight relation between the MoAs of a compound and its therapeutic indications. This last section will exemplify how actual drug repurposing hypotheses can be derived from the FTC. The method presented here makes use of the FTC categories as compartment of a toolbox helping to find drugs to fix Alzheimer’s disease. The categories of the classification are indeed representing the roles of compounds in the human body and they can be seen as many compartments for chemical tools with special biological purposes. Five FTC categories containing drugs are related to the neurodegenerative condition: *Anti-amyloid precursor protein biosynthetic process agent* (FTC\_A0042983), *Anti-tau-protein kinase activity agent* (FTC\_A0050321), *Anti-tau protein binding agent* (FTC\_A0048156), *Anti-beta-amyloid binding agent* (FTC\_A0001540) and *Pro-beta-amyloid binding agent* (FTC\_P0001540). Note that only processes and functions directly related to the disease have been cherry picked, the idea is to find drug candidates that could somehow outrightly influence the condition. We have then considered the drugs present inside each of these classes as many potential candidate. Figure 4 shows these drugs, which have been further manually grouped based on the overall similarity of their actions (letters on Figure 4).

Figure 4: FTC categories (letters A, F, H, I and J) describing some of the modes and mechanisms of action that could impact Alzheimer’s disease. Drugs classified in these FTC categories are listed and further manually grouped based on their mode of action similarities (subgroups B, C, D, E and G).

The subgroups B, C and D are inhibitors of the cholinergic system and some of them, such as the galantamine (DB00674), are already investigated to treat Alzheimer’s disease and other related dementia (pubchem). This class of agent is in line with the *cholinergic hypothesis* (cite PMID:10071091), stating that a Alzheimer’s disease could be caused by dysfunctions in the processing of the acetylcholine. The subgroup E is exclusively composed of barbiturates (central nervous system depressants). The presence of this pharmacological class of compounds as an Alzheimer treatment is more surprising, as very few literature reports on it. Further investigations reveal that the neuronal acetylcholine receptor subunit alpha-7 (P36544), a common off-target of barbiturates, binds the beta-amyloids with high affinity. As the beta-amyloids are themselves strongly involves in the pathology (cite <http://europepmc.org/abstract/MED/10681545>), barbiturates could affect the state of the condition, similarly to cholinergic inhibitors.

The group F contains four compounds; the nicotine and the varenicline have been grouped because of similar pharmacology (group G). The nicotine has been shown to improve some of the symptoms of Alzheimer’s disease (PMID:1410164), it is therefore expected to find this molecule in the predictions. The varenicline possess a pharmacology related to the nicotine, which would explain the presence of the drug in this category. The two remaining drugs of group F are respectively the pralidoxime and the dipivefrin; few information is available regarding their potential action against the disease, yet these compounds seem linked the cholinergic hypothesis and could be considered for experimental testing.

The groups H and I contain respectively one molecule each. These compounds have been classified as agents perturbing some of the physiological function of the Tau protein, key actor in Alzheimer's disease (cite). The vorinostat (group H) is currently indicated for *the treatment of cutaneous manifestations in patients with T-cell lymphoma* (drugbank), yet a study shown *in vivo* (mouse model) the potential of the drug and other histone deacetylase inhibitors in regards to memory defficit (cite <http://www.nature.com/npp/journal/v35/n4/full/npp2009197a.html>). The presence of the lithium (group I) was unexpected and suspected of being a false positive at first glance. However a recent study demonstrated a long-term protective effect for the ion in regards to Alzheimer’s disease (PMID:22500970 and 21525519). The last group (J) contains the ezetimibe and the hesperetin. These two compounds are primarily used as cholesterol lowering agents (statins). As the cholesterol metabolism in the brain appears to be related to the dementia, the statins are believed to prevent or improve the symptoms of the patients. Early studies (PMID:17877925) are however failing to clearly show a beneficial effect, yet the investigation is still open.

Out the examples briefly presented above, reported and confirmed by the literature, the FTC appears to be suitable to identify real repurposing hypotheses tailored to a disease. The hypotheses presented before appeared to be already known, yet it validates the potential of the resource. From the choice of wanted MoAs (analogous to compartments in a toolbox), the user can retrieve the compounds which might impact the treatment of a given set of symptoms.

# Discussion

The FTC is a novel classification for approved drugs which can be used as a starting point to generate drug repurposing hypotheses. We have guided the reader through the various ways the resource can be explored. No experimental validations have been presented in this manuscript, yet references to supporting literature was provided in order to legitimate the claims. Our main goal with this document was to introduce the resource to the community; deeper analyses are performed over the classification at the time of writing. We will now discuss the interpretation of results and hypotheses generated out of the classification.

## Biological assumptions

A asset of the FTC is its ability to handle efficiently categorical data: Classes and relationships are finely defined, in order to classify compounds based on the semantics of their relations. The properties linking drugs to their respective protein targets (*positive* and *negative perturbations*) are however simplistic. At the time being, no consideration is given regarding the binding strength between the drug and the proteins, yet it is a key factor to derive potent and specific activities in the human body. This consideration relates to other types of numerical data, such as the dosage; the FTC can predict a role for a drug, yet without providing any more information about the concentration or the administration route necessary to obtain the potential effects. The current relations between targets and their involvement in biological processes are also not a fully accurate representation of the biological phenomenon. In a cell, specific domains of the protein could mediate different functions. Only one of such activity type can sometimes be inhibited by a drug, yet we are assuming in the FTC that as long as a drug affects a protein, it can therefore alter all its known functions.

These limitations are coming from the semantics behind the axioms structuring the classification, themselves based on the information available from the databases (e.g. GOA, DrugBank, etc...). Despite entailing more information than the biological reality conveys, the axioms help to generate a larger number of hypotheses, primary goal of the FTC. The dosage issue is partially addressed by the *regulator pattern* (see Material and Methods): It should be easier to experimentally adjust the concentration of the compounds classified as *pro* or *anti* biological process agents in order to modulate a physiological effect.

The predictions generated by the FTC depend on the resolution of the curated information released by the original data providers too. Erroneous or missing information will lead to misclassification by the reasoner for instance. Some expected outcomes are also missing from the predictions; the sildenafil for instance was expected to be classified as *anti-penile erection agent* (FTC\_A0043084), yet the lack of appropriate GO annotation prevents it. After discussion with the GOA curation team, a manual annotation can only be asserted based on published experimental results. No document was found to support the involvement of the cGMP-specific 3',5'-cyclic phosphodiesterase (O76074 - sildenafil’s main target) in the negative regulation of penile erection (GO:0060407), therefore no annotation can be made. Further work could be done in this direction, by trying to automatically infer more annotations or by using the electronically generated ones, in order to generate broader however potentially fuzzier repurposing hypotheses.

## Interpreting the evaluation

It is possible to evaluate the work against other resources, containing classes describing similar concepts than the ones present in the FTC. Here we have assessed the content of the FTC against the ATC, knowing that these two taxonomies have diverging goals. During the evaluation, equivalences have been manually asserted between classes which are assumed to have fairly similar meaning and containing similar sets of compounds. These manual assertions are however a weakness, as they are themselves not evaluated (free parameter). The presence or absence of a link was determined only by one curator and any mistake can influence consequently the recall and precision values. In this regards, the evaluation should be considered more as a safety control rather than a formal assessment of a predictive method. What was considered as a gold standard (the ATC) is in practice not a true one and the evaluation should be carefully interpreted accordingly. The predictions directly derived from the evaluation should be interpreted with the same caution, as it is currently impossible to isolate a false positive from reprofiling opportunity. These considerations do not interfere with the repurposing predictions generated based on semantic similarities or discrete categories as presented in the Exploration section. Finally, note that the ATC/FTC equivalences are open and editable online, any modification will be automatically incorporated in the next release of the resource. It is also possible to evaluate the FTC against a different taxonomy, like the Medical Subject Headings for example, which can be subject to future work.

## Conclusion

This manuscript introduced the FTC, a free, open and extensible resource which can assist drug repurposing initiatives or enhance computational studies making use of the concept of *mode of action*. The construction of the classification relies on description logics as a mean to define the mode and mechanism of action of approved drugs. The work leverages the semantics of distinct databases, funded and working in parallel on different thematic. The platform will be further used to generate predictions in a systematic fashion, which would then be experimentally tested in the laboratory for validation.

# Materials and Methods

The aim of the FTC is to first provide a list of categories describing the mode and mechanism of action of drugs. Then in a second step the newly created categories are automatically populated with approved compounds. Finally, the FTC is evaluated and repurposing hypotheses can be generated.

## Availability and implementation

The code behind the creation of the resource is open and available at <https://github.com/loopasam/ftc>. The web application built on the top of the FTC can be find at <https://www.ebi.ac.uk/chembl/ftc> and the documentation can be accessed at <https://github.com/loopasam/ftc/wiki>. The reader should be familiar with description logics and the Web Ontology Language (OWL) to fully understand the construction of the knowledge base. An introduction to description logics from the perspective of the biomedical scientist is available on the wiki at <https://github.com/loopasam/ftc/wiki/Description-Logics>. The FTC implementation relies mostly of Brain (cite) and the web application builds on the top of the Play! framework (cite). Classification tasks use the ELK reasoner (cite). The computer hosting the web application has 8Gb of memory with 4 processors, this architecture allows to unleash fast parallel reasoning, thanks to ELK’s design. More functionalities will be added to the web application following user’s requirement (lean implementation).

## FTC categories creation

The mode of action categories present in the FTC are defined based on the terms coming from the Gene Ontology (GO). Both the molecular function and biological process branches are used for this purpose, yet handled differently.

### Categories related to biological processes

All the biological processes featured by the GO are looked-up one by one. All the time a process is linked to another process (*X*) via a positive or negative regulation link, two FTC classes are created: *Anti-X agent* and *Pro-X agent*. For instance the GO term *positive regulation of blood coagulation* is linked to the term *blood coagulation* via a *positively regulates* relation, therefore two FTC categories *Anti-blood coagulation agent* and *Pro-blood coagulation agent* are created. The identifiers of the new FTC classes are also derived from the GO term used to create the class pattern. The GO numeric identifier is re-used and the letter *A* or *P* is appended before to emphasize the *anti* or *pro* pattern. From the example presented previously, the FTC class *Anti-blood coagulation* has FTC\_A0007596 as identifier, because the GO term *blood coagulation* is referenced by GO:0007596. Following the same logic, FTC\_P0007596 is the identifier of the class *Pro-blood coagulation*. The design choice for identifiers and labels allows to fully rely on the high quality work provided by the GO curation team and scale over it.

### Categories related to molecular functions

The mode and mechanism of actions related to molecular functions are created in the following manner: All the time a molecular function (*Y*) is encountered then two FTC categories are created, as with processes: *Anti-Y agent* and *Pro-Y agent*. The identifiers are assigned the same way as described before. For instance, out of the Gene Ontology term *androgen receptor activity* (GO:0004882) two FTC classes are derived: *Pro-androgen receptor activity agent* (FTC\_P0004882) and *Anti-androgen receptor activity agent* (FTC\_A0004882).

## Equivalent definitions

FTC classes are generated as presented in the previous section. Up to this point, these categories are only individual tokens with a human readable label as well as an identifier. The next step is going to assign equivalent definitions to each FTC class. A reasoner can understand such definitions and will automatically classify the knowledge base accordingly; drugs will then be assigned to FTC categories and the taxonomic structure arises after the reasoning step. Equivalent definitions are written as OWL class expressions using the entities of the knowledge base. The core classes and properties are presented in the supplementary material, this section assumes that the reader is already familiar with the FTC specifications (summarised at <https://github.com/loopasam/ftc/wiki/Knowledge-Base>). There are two types of equivalences: The first one captures perturbation of regulatory biological processes (so called *regulatory patterns*) and the second one handles the perturbed functions (*functional patterns*).

### Regulatory pattern

Some of the FTC categories are created from the biological processes present in the GO (cf previous sections); these categories have two arbitrary equivalent definitions, representing the two possible ways a compound might impact the biological process. *Anti-biological process agent* FTC categories contain the drugs that negatively perturb a target involved in the positive regulation of the biological process. The *anti* categories also feature the compounds that positively perturb a negative regulator of the same process. The *pro* categories are equivalent to the opposite pattern. Figure 5 illustrates the equivalent definitions for the FTC class *Anti-blood coagulation agent* (FTC\_A0007596).

Figure 5: Example of equivalent definitions (gray boxes) for the FTC category *Anti-blood coagulation agent* (FTC\_A0007596). After the data integration step, a reasoner will look the definitions and identify drugs that satisfies them. Definitions are expressed using the Web Ontology Language (OWL), serialised here using the Manchester syntax.

### Functional pattern

The FTC categories generated from the GO molecular functions are also equivalent to a logical definition. *Anti* FTC categories dealing with molecular activities are asserted as equals to the drugs that negatively perturb a the function. *Pro* categories are equivalent to the drugs that positively perturbs the function of interest. A summary of the patterns definitions in available on the wiki online at <https://github.com/loopasam/ftc/wiki/Mode-of-Action>.

## Data integration

At this stage, the knowledge base contains the created FTC classes associated with their logical definitions, as well as the GO and the core FTC entities. The knowledge base is then further populated with some information coming from various public databases. Only manually curated information is considered. This section illustrates how the core entities interact with the different data types in order to create the axioms.

### DrugBank

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information (cite). The approved drugs (small molecules and biotech) acting on proteins are extracted from the database and imported in the FTC knowledge base. In order to be selected, a compound must firstly be approved and secondly have an action on at least one human protein target present inside Uniprot. The protein targets all have at least one manually asserted GO annotation for a biological process or a molecular function. DrugBank links compounds to targets via *actions*. The DrugBank actions are somehow structured and consistent: Concepts such as *inhibitor* or *agonist* are reused throughout the database for example, yet they are not strictly formalised as a controlled vocabulary. These actions are manually standardised to the core properties of the FTC according to their meaning: For instance the action *antagonist* is mapped to the FTC *negatively-perturbs*.

Compounds coming from DrugBank are represented as OWL classes and asserted as subclasses of *DrugBank compound* (FTC\_C2). Protein targets are described as classes too and subclasses of the core class *protein*. Each DrugBank compound is then connected to its target via the following axiom pattern: *drug SubClassOf perturbs some protein*. E.g. *Ximelagatran SubClassOf negatively-perturbs some Prothrombin*.

### Gene Ontology annotations

The GO annotation program aims to provide high-quality GO annotations to proteins inside UniProt (cite). In the context of the FTC, such annotations are used to create axioms linking protein targets to molecular functions and biological processes. Each protein annotated with a function creates an axiom such as *protein SubClassOf has-function some molecular function*. Each protein annotated to a biological process creates an axiom such as *protein SubClassOf involved-in some biological process*. E.g. *Prothrombin SubClassOf involved-in some positive regulation of blood coagulation*. Each protein can be involved in multiple processes and capable of realizing multiple functions.

## Knowledge base classification

The knowledge base is fully built at this step and contains core classes, mode of actions descriptions alongside the actions of approved DrugBank compounds on Uniprot targets. The proteins are linked to their molecular functions and involvement in biological processes via the GO annotations. The logical specifications of the FTC are there to glue the different data together and to explicitly express the logical links between resources. The FTC knowledge base follows an OWL2 EL profile, which enable the use of fast and parallelised reasoners such as ELK (cite). During the classification process, the reasoner checks whether the mode of action equivalent definitions are satisfied or not and assigns drugs inside the corresponding FTC categories. The tree structure of FTC appears also at this step.

## Evaluation methodology

As the classification of therapeutic agents is done in an automated way, it is important to evaluate the results generated against a known resource which will be considered as gold standard. The assessment of the FTC is done against another similar classification, the Anatomical Therapeutic Chemical Classification System (ATC) (cite). The ATC has been developed to serve as a tool for drug utilization research in order to improve quality of drug use. The goal of the ATC differs from the one of the FTC, yet the two resources are sharing some very similar concepts which can be used for the evaluation. Categories of both classifications contain approved drugs with a DrugBank identifier, meaning that some of the drugs indexed in the FTC are also present in the ATC. From that, it is possible to define some evaluation points, which will help to assess the automated classification process.

### Evaluation Points

An evaluation point is defined as an equivalence between a class from the FTC with one or more classes from the ATC. The idea is to look at the set of drugs contained in both side of the equivalence and derive a number representing the overlap out of it. This is illustrated in the Figure 6. Evaluation points are defined by hand and not themselves evaluated. The full list of evaluation points as well as a summary of the results are available online at <https://www.ebi.ac.uk/chembl/ftc/evaluation/>. Each evaluation point has a series of true/false positive and false negative drugs associated with it.

Figure 6: Example of evaluation point. Classes from the ATC are manually mapped to classes from the FTC. The drugs present in each side of the equivalences are compared, and some metrics are derived from it (false positive/negative and true positive). A summary of the evaluation is available online at <https://www.ebi.ac.uk/chembl/ftc/evaluation/>.

### True Positives

Drugs that are present in both the FTC and the equivalent ATC class(es) are called true positives. These compounds reflect that the automated classification was capable of retrieving correctly the information present in the gold standard (ATC).

### False Negatives

These drugs are present in the ATC class(es) but not in the corresponding FTC class. The automated classification failed to retrieve these compounds. The smaller the number of false negatives is, the better the FTC is at recalling drugs. A small number of false negatives means that if a drug is present in the ATC (gold standard), then it is likely that this drug will also be correctly categorised in the FTC.

### False Positives

The false positives are the drugs present in the FTC category of the evaluation point but not in the corresponding ATC classes. A high number of false positives means that the FTC is over-assigning compounds to classes. The false positives relates to the accuracy of the classification. In the context of this work, some false positives could also be considered as drug repurposing opportunities.

### Precision

Precision is the probability that a randomly selected drug from the FTC is present in the equivalent ATC classes. The value is standardised as a percentage and corresponds to the formula: True Positive / (True Positive + False Positive).

### Recall

Recall is the probability that a randomly selected drug from the ATC has been assigned to the correct corresponding class in the FTC. The value is standardised as a percentage and corresponds to the formula: True Positive / (True Positive + False Negative).

## Semantic similarity

The semantic similarity measure performed over the FTC is a derivative of the Jaccard index (cite). It is probably best understood as an example: If we consider two classes A and B, the semantic similarity between these classes corresponds to the number of OWL superclasses (direct and indirect, obtained with a reasoner) that are shared by A and B (intersection) divided by the number of superclasses of A or B (union). The index ranges from 0 (totally different) to 1 (identical).

## Mode of action similarity against indication

A statistical analysis was performed over the data presented on Figure 3. When two compounds are randomly taken, they have on average a higher mode of action similarity when they are assigned to the same ATC category (one ATC level). This observation can be observed on Figure 3; squares shape around the central diagonal. In order to estimate whether this observation was due to chance only, we formulated the following null hypothesis: *The pairs of drugs belonging to the same ATC category have on average the same MoA’s similarity when compared to the compounds belonging to different ATC categories*. A permutation test was performed (n = 20’000), the null hypothesis being rejected for a significance level below 0.05. The choice for a permutation test was driven by the fact that MoA similarity values do not follow any type of standard distribution (data not shown). For the significance level set, the null hypothesis was rejected for all first level ATC categories.

## Knowledge base specification - Supplementary Material

This section will introduce the reader to the scaffold of the knowledge base underlying the FTC. The logic structuring the FTC comes essentially from a set of core OWL properties (rich RBox). Some of these properties originate from the GO. When necessary some new ones have also been introduced. In order to understand how these properties interact, first will be presented the fundamental classes present at the top of the FTC classification.

### Core classes

The high level concepts covered by the FTC are represented as OWL classes and enumerated below. Some of these core classes are coming from external ontologies, in which case the original URI is preserved.

**molecular function**

Identifier: http://purl.obolibrary.org/obo/GO\_0003674

Label: molecular function

Definition: As defined by the Gene Ontology. Elemental activities, such as catalysis or binding, describing the actions of a gene product at the molecular level. A given gene product may exhibit one or more molecular functions.

**biological process**

Identifier: http://purl.obolibrary.org/obo/GO\_0008150

Label: biological process

Definition: As defined by the Gene Ontology: Any process specifically pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. A process is a collection of molecular events with a defined beginning and end.

**Protein**

Identifier: http://purl.uniprot.org/core/Protein

Label: protein

Definition: As defined by Uniprot. Description of a protein.

Comment: Gene products present inside the FTC are all human proteins. Uniprot URIs are used.

**Drug**

Identifier: http://schema.org/Drug

Label: drug

Definition: As defined by schema.org. A chemical or biologic substance, used as a medical therapy, that has a physiological effect on an organism.

Comment: In the context of the FTC, DrugBank chemicals are considered for their role as therapeutic agent rather than for their chemical structure. Mode of action classes are representing agent types and are subclasses of this concept.

**therapeutic agent**

Identifier: https://www.ebi.ac.uk/chembl/ftc/FTC\_C1

Label: therapeutic agent

Definition: Role of a drug capable of producing a therapeutic effect.

**DrugBank compound**

Identifier: https://www.ebi.ac.uk/chembl/ftc/FTC\_C2

Label: DrugBank compound

Definition: Drug coming from DrugBank.

### Core properties

The expressivity of the FTC comes mostly for the properties (RBox). Below is a list of the OWL object properties present inside the FTC knowledge base.

**part-of**

Identifier: http://purl.obolibrary.org/obo/BFO\_0000050

Characteristic: Transitive

Label: part-of

Definition: As defined and used in the Gene Ontology. More information at <http://www.geneontology.org/GO.ontology.relations.shtml#partof>

**has-part**

Identifier: http://purl.obolibrary.org/obo/BFO\_0000051

Characteristic: Transitive

Label: has-part

Definition: As defined and used in the Gene Ontology. More information at <http://www.geneontology.org/GO.ontology-ext.relations.shtml#haspart>

**regulates**

Identifier: http://purl.obolibrary.org/obo/RO\_0002211

Chained property: regulates o part-of -> regulates

Label: regulates

Definition: As defined and used in the Gene Ontology. More information at <http://www.geneontology.org/GO.ontology.relations.shtml#regulates>

**negatively-regulates**

Identifier: http://purl.obolibrary.org/obo/RO\_0002212

SubPropertyOf: regulates

Label: negatively-regulates

Definition: As defined and used in the Gene Ontology. More information at <http://www.geneontology.org/GO.ontology.relations.shtml#regulates>

**positively-regulates**

Identifier: http://purl.obolibrary.org/obo/RO\_0002213

SubPropertyOf: regulates

Label: positively-regulates

Definition: As defined and used in the Gene Ontology. More information at <http://www.geneontology.org/GO.ontology.relations.shtml#regulates>

**involved-in**

Identifier: https://www.ebi.ac.uk/chembl/ftc/FTC\_R1

Label: involved-in

Domain: protein

Range: biological process

Definition: Entails the participation of a protein in a biological process

**has-function**

Identifier: https://www.ebi.ac.uk/chembl/ftc/FTC\_R2

Label: has-function

Domain: protein

Range: molecular function

Definition: Describes the molecular function a protein can realize.

**perturbs**

Identifier: https://www.ebi.ac.uk/chembl/ftc/FTC\_R3

Label: perturbs

Domain: drug

Range: protein

Definition: Specific biochemical interaction through which a drug substance will affect the activity of a protein (mechanism of action). The property refers to the specific molecular targets to which the drug binds, such as an enzyme or receptor.

**negatively-perturbs**

Identifier: https://www.ebi.ac.uk/chembl/ftc/FTC\_R4

Label: negatively-perturbs

SubPropertyOf: perturbs

Definition: Specific biochemical interaction through which a drug substance will decrease the activity of a protein. The property refers to the specific molecular targets to which the drug binds, such as an enzyme or receptor.

**positively-perturbs**

Identifier: https://www.ebi.ac.uk/chembl/ftc/FTC\_R5

Label: positively-perturbs

SubPropertyOf: perturbs

Definition: Specific biochemical interaction through which a drug substance will increase the activity of a protein. The property refers to the specific molecular targets to which the drug binds, such as an enzyme or receptor.