Executive summary (not exceeding 500 words)

IIT Kharagpur (IIT KGP) a public engineering institution established by the government of India in 1951 has entered into a strategic collaboration with (GVK BIO), Hyderabad, one of Asia's leading Discovery Research and Development organizations. The Collaboration is focused on advancing the technologies in the field of Drug Repurposing (DR) a rapidly evolving process to find new therapeutic uses for pre-existing and failed drugs and Drug Reconsidering (DRe), a form of stratified medicine for categorizing patients based on disease risk, dosage predictions or therapeutic response. The joint proposal aims at accelerating Drug Repurposing and Drug Reconsidering research.

The proposal envisages on harnessing the skill set and technological advancements of each of the organisations. As part of the project, the joint team of IIT KGP and GVK BIO, will work on Natural Language Processing (NLP) and Machine Learning (ML) algorithms to extract entities, identify co-references and semantic relationships between entities from biomedical literature.

NLP will be used for gathering biologically meaningful information from clinical trials registries by integrating several meta-layers of textual information into standard information models. Text mining approaches will be used for associating pharmacological data to drugs and biomarker extraction from literature. Sequence learning algorithms including deep neural network architectures and conditional random fields will be used for this task.

To enable fast progress, the joint team, will use publicly available ontologies and GVK BIO databases, along with the manually curated Drug-Disease-Target relationships data and pharmacological data made available by GVK BIO, for training machine learning algorithms. Focus will be on data integration and validation algorithms, for combining and validating relations and associations from various sources.

Systems biology based approaches will be employed to understand disease pathophysiology and predict possible mode of treatment. A comprehensive diseases-protein network would be built utilizing integrated data from GVK BIO Drug Repurposing Integrated Database (GRID, an extensive database of GVK BIO) and through literature mining. Text Mining will be used for finding edge directions and semantic types of relations and finding kinetic parameters). Network generation and Simulations will be performed to integrate disease specific gene expression data.

The program to a major extent will accelerate the technologies for DR/DRe research and significantly cut down on costs and bring in the much needed efficiency and effectiveness of application of DR/DRe approaches, which in turn will be commercialised by GVK BIO.

Background and motivation (not exceeding 500 words)

The process of development of high quality drugs with fewer side effects is lengthy, costly, and has high attrition rates. Drugs fail due to inefficacy and safety reasons. Drug Repurposing minimizes clinical trial failures, thereby reducing fiscal losses and expediting drug approvals. The global market for drug repurposing reached nearly $24.4 billion in 2015 and should reach $31.3 billion in 2020, reflecting a five-year compound annual growth rate (CAGR) of 5.1%. Drug failure may be caused by incomplete understanding of diseases and genetic variability of patients. Traditional methods for DR based on phenotypic drug-screening and serendipitous clinical studies observations are time consuming and expensive.

Computational approaches overcome these shortcomings, and are multi-dimensional, incorporating drug and target structural data, drug efficacy, adverse events information, signalling pathways, interaction networks, genomewide association studies, gene expression data and clinical trial outcomes. These approaches require integration of publicly available data which is heterogeneous. Network based methods utilize OMICS data, pathway information and protein-protein interactions. Mathematical models of human systems enable dynamic simulation of different disease related parameters. Stratification of patients based on genotype enhances the chance of success for drug candidates and provides indications for personalized medicine.

Computational approaches have been further strengthened by biomedical NLP of published literature to generate novel testable hypothesis from indirect associations in literature.

GVK BIO uses GVK BIO Repurposing Integrated Platform (GRIP) and Database (GRID) for arriving at novel hypothses for DR using drug-, target-and disease-centric approaches. GRIP consists of a customized Drug Repurposing Database, Proprietary repurposing algorithms, internally developed analytical engine and visualization tool. GRIP centered on drug, target and disease associations, has all required components to address the repurposing/repositioning activity and utilizes validated multipronged approaches to identify new indications for drugs/compounds repurposing.

GRID is a custom made proprietary database which includes data from both public and proprietary sources. It contains individual data marts of drugs, diseases and targets and also creates multi-dimensional profiles of biologically relevant entities such as genes, pathways, biomarkers and adverse events.

Currently GVK BIO relies on public databases and manual curation for generating drug-target-diseases associations. The processes involved are time taking and incomplete, and can be markedly accelerated by using automated text mining approaches on published literature. Besides invaluable insights can be gathered from other unstructured resources like Clinical trials and patents. Employing NLP can potentially increase the usability of such unstructured resources which are currently under-represented in Drug Repurposing.

Project outcomes (please list specific objectives): *The project should address a specific need of the industry/industries and there should be clear expected outcomes from the project. It is expected that joint patents will result from this project.*

1. Accelerating DR research: The extensive and accurate drug-disease-target relationship knowledge base will aid in quicker and precise prediction of drug repurposing activities, which in turn would enhance productivity/risk ratio.
2. Developing new algorithms and knowledge base for computational repurposing and reconsidering. Physiological network simulations, prediction of drug toxicity and adverse reaction through network analysis and development of Machine Learning approaches to predict patient response to drugs.

Scope (not exceeding 1500 words): *The scope should clearly lay out the contributions of the academic partner and the industry partner.*

A. Data integration: GRID is a customized proprietary database containing compilations from over 40 public databases other than in-house curated data. GVK BIO Scientists rely on GRID for identification of novel co-relations between drugs, diseases and targets. Continued up-gradation of GRID is essential for highest accuracy of prediction and on-par performance with competitors. Till date GRID has mostly relied on manual curation of data from published literature and public domain databases. Although recent activities have included automated approaches to some extent, the proposed project aims to include the benefit of NLP to update GRID through semantic data integration. NLP will be extensively used to extract relevant information from biomedical literature (PubMed) and other web-platforms to upgrade GRID on multiple layers, viz.,

i) drug structural information,
ii) protein-protein interactions network,
iii) co-relation of gene expression patterns with genomic or other variations etc.

Considering the huge volume and varieties of biological data, heterogeneity and redundancy are the two major problems associated with seamless integration of data. In-built capabilities of NLP would be ideally suitable to nullify their effects. Along with semantic web ontologies, a custom training set will be available to create better performing Bio-NLP, which would be used to incorporate more data in to GRID.

B. NLP techniques on PubMed and patents: GVK BIO during the course of its business has developed text-mining based Drug, Target and Disease relationships, searching in biomedical literature. Text-mining based co-occurrence statistics provides a reasonable number of relationships with good signal to noise ratio. At GVK BIO, these initial relationships are manually curated by biologists to provide mechanism of action based drug repurposing leads. A database of such manually curated drug, target, and disease relationships exists with GVK BIO. For automating this exercise, two tasks are important.

1) Extraction of drug, target and disease names along with synonyms.
2) Semantic relationships between the named entities using NLP methods.

For this task, several open-source ontologies exist which can be improved to find relationships more efficiently.

IIT KGP has expertise in NLP, Information Extraction and Machine Learning (ML). The joint team will work on GVK BIO curated data, various databases and ontologies as the knowledge base. The syntactic and discourse relationships in the corpus and the knowledge base will be used for obtaining features. Sequence learning and deep neural network based architectures, vector representations of entities and their attributes, will be explored for automating the above tasks.

Another important and as yet unaddressed problem is associating pharmacological data to drugs. A few methods explored lacked a comprehensive training dataset. GVK BIO has manually curated database of pharmacological data that can be utilized as training data for this exercise, and there is a scope for using sophisticated machine learning techniques for this task.