# Implementation of ChemBERTa for Chemical Fingerprinting of Molecules and Property Prediction at BPC INT'L

**(Work Plan & Project Initiation (P.I.) Sheet)**

| | | | |
|---|---|---|---|
| **PROJECT TITLE:** | **Implementation of ChemBERTa NLP Algorithm for Chemical Fingerprinting of Molecules and Property Prediction at BPC INT'L** | **PROJECT NUMBER:** | |
| **PROJECT CONSULTANT:** | Primary Project Consultant & other Consultants from ABC_&_Co who possibly had worked with BPC INT'L Int'l in previous projects. | **DATE INITIATED:** | 00/00/0000 |
| **KEY CUSTOMERS:** | Big Pharma Company Int'l | | |
| **PROJECT SPONSOR:** | Big Pharma Company Int'l | | |

**OBJECTIVE:** *What is the purpose of the project?*

Chemical fingerprinting is used to identify and characterize a substance by analyzing its unique chemical composition. This process essentially creates a "fingerprint" of the substance, thus allowing for comparison to known samples. Chemical fingerprinting enables tasks like materials quality control, source identification, and contaminant detection in various fields like environmental monitoring, food and drug analysis, forensics, molecules analysis and drug discovery to be easily accomplished.

ChemBERTa, an adaptation of the popular RoBERTa, machine learning and natural language processing (NLP) algorithm has been found to be quite useful for chemical fingerprinting and molecules properties prediction when trained on the Simplified Molecular Input Line Entry System (SMILES) data set. SMILES is the world largest open sources collections of chemical and molecules structures.

Big Pharma Company Int'l is seeking to leverage its extensive materials data sets and the ChemBERTa algorithm for chemical fingerprinting, molecules identification and representation, and molecules properties prediction.

ABC_&_Co engineers will be able to assist BPC INT'L by leveraging ABC_&_Co's expertise in data engineering, machine learning, software development and natural language programming processing (NLP) to achieve an implementation of the ChemBERTa algorithm using BPC INT'L data.

Refine objectives by further discussion with project key customers

**DELIVERABLES:** *What are the major deliverables or focus of the project?*

a. A ChemBERTa based NLP model that can do molecules property prediction using BPC INT'Ls data.
b. Integration of the developed model for downstream applications at BPC INT'L.
c. Integration will be accomplished by deploying the model for BPC INT'L Int'l data property prediction through cloud and API (Application Programming Interface) end point deployments.
d. Reliability of the deployed model will be accomplished by deploying through Docker containerization and Kubernetes deployments.

Refine deliverables by further discussions with key project customers

**SCOPE:**

Covered:
BPC INT'L Intl data curation, data cleaning, and extraction, transform and loading processes.

# Implementation of ChemBERTa for Chemical Fingerprinting of Molecules and Property Prediction at BPC INT'L

**(Work Plan & Project Initiation (P.I.) Sheet)**

ChemBERTa algorithm-based prediction model that can be used to achieve chemical fingerprinting and molecules property prediction of materials.

Integration of the model for day to day use at BPC INT'L

Not Covered:
To be discussed with key customers

| **JUSTIFICATION / BENEFITS:** | *Why initiate project? What are business reasons / paybacks for doing project? (Impact, **C**ost **o**f **P**oor **Q**uality [COPQ], etc.,* |
|---|---|

(1) Enhanced molecules prediction, material identification and chemical fingerprinting
(2) Better and faster drug discovery
(3) Better and faster understanding of the biological, chemical and physical nature of materials through their chemical fingerprints
(4) Possibilities of developing targeted drugs for specific BPC INT'L customers, leading to better customer engagements, customer retention and opportunities for new businesses and applications.
(5) If BPC INT'L does not develop the model on time and at cost, there is possibility of their losing vital chunks of their customers to local and international rivals

| **RISKS:** | *What potential problems are associated with the project?* |
|---|---|

- Limitations of existing dataset – Small dataset of some material can lead to low prediction accuracy – Possibly imbalanced data: special treatment required.
- It is expected that BPC INT'L will have a big structured and unstructured data set. Curating and data selection can take a lot of time.
- Pharmaceutical data set due to the possible connection with human data set through usage always need special care due to government regulations and security.

| **ASSUMPTIONS:** | *What is required for the project's success? (i.e.: resources, budget, development time, etc.)* |
|---|---|

Deloitte ABC_&_Co resource time: Estimated 245-400 hours total. This can translate to ~ $72,000 if ABC_&_Co engineers charges for their time at the rate of $180/hr. Other miscellaneous charges from ABC_&_Co can be added as appropriate.

| **RESOURCES / ROLES:** | Include required resources / team members / responsibilities on the project |
|---|---|

Main Resources:

- Project Manager (Data Engineering / AI Applications)
- 2 Data Engineer with knowledge of GPU, RAPIDS (cuDF, Dask), Python Vaex, and/or OpenCL libraries

Responsibilities:
Curating and transforming (tokenization, vectorization) BGC's data. These engineers must have adequate experience with traditional ETL applications and cloud-based ELT applications.

BGC already have cloud capabilities, hence ABC_&_Co engineers shall work along with BGC team to ensure that relevant data (original and transformed data) are available for the downstream NLP engineers.

The Data Engineers should be able to work with both relational and NoSQL databases since it is expected that clients such as BPC INT'L will have very big structured and

unstructured data sets. Other responsibilities, including cleaning the data, and possibility separating the data from 3D molecular data to 1D and 2D data sets as needed.

- 2 NLP Engineers with knowledge of HuggingFace suite of libraries.
  Responsibilities:
  Model training through transfer learning. Use BPC INT'Ls data to fine-tune ChemBERTa. Based on availability of time and resources, it may be desirable to implement and train RoBERTa algorithm from ground up since the authors of ChemBERTa mentioned that prediction performance of ChemBERTa could be improved by training over more epochs.

- Data Visualization Engineer with knowledge of 2D and 3D material visualizations.
  Responsibilities:
  Molecules exists with 3D geometries (viz: biological, chemical and physical properties). Hence, at least 1 data visualization engineer with extensive knowledge of open-source 3D graphics tools such as Plotly (with dash bio), ChemPlot, RDKit, Mayavi, PyVista etc.
  The visualization engineer is also expected to be comfortable with rendering results on dashboards.

- 2 molecules/chemical/material engineers and/or domain experts from BPC INT'L.
  Responsibilities:
  These domain experts shall serve as initial sounding board to advise the data and machine learning engineers regarding the suitability of their model to BPC INT'L's test data sets.

- 2 BPC INT'L Engineers:
  Responsibilities:
  For future model retraining and prevention of model drift, it is useful to have at least 2 engineers (data and NLP engineer) from BPC INT'L to work along with engineers from ABC_&_Co. It will be great if these BPC INT'L engineers have working knowledge of Lambda architecture so that they can be able to effect retraining of the developed model using newly available batch and if required, to use streaming data in real time as well. These engineers can always reach out to Deloitte ABC_&_Co consultants if they have issues that are not within their knowledge domain.

- Software Engineer:
  Responsibilities:
  To transition the model into a deployable software. BPC INT'L may need the software with different
  types of access points for many of their molecule/chemical fingerprinting engineers. They may also need to a desktop version of the software. The software engineer will work
  along with the DevOps Engineer to package the software in the needed formats. For example, desktop deployment, deployment through Kubernetes cluster for reliability and software availability.

- DevOp Engineer:
  Responsibilities:
  Work along with the Software Engineer to accomplish dockerization, Kubernetes deployment and provisioning for desktop version of the software.

# Implementation of ChemBERTa for Chemical Fingerprinting of Molecules and Property Prediction at BPC INT'L

**(Work Plan & Project Initiation (P.I.) Sheet)**

Consultants:

- Emmanuel Oyekanlu (Data/AI/Softwarization), other consultants from ABC_&_Co.

**Hardware Needs:**

For this project, it could be reasonably assumed that BPC INT'L shall provide the ABC_&_Co team access to their cloud platform. Hence the cost of using BPC INT'L's cloud platform for all phases of the project shall be paid by BPC INT'L.

Hardware instances for this type of project always have needs for parallel processing of data. If BPC INT'L cloud instances does not yet have support for GPU based parallel processing, request shall be made for cloud instances that support parallel processing capabilities. It was reported that an NVIDIA V100 GPU hardware was able to train 10million SMILES data within 48 hours.

**Software Needs:**

Many of the software tool that will be needed for this project may be readily available as part of the BPC INT'L cloud platform. Since BPC INT'L have a competent cloud team, it is reasonable to assume that BPC INT'L's cloud platform may have necessary parallel processing cloud software instances such as PySpark, and RAPIDS suite of libraries.

BPC INT'L can also be urged to make subscription for Databricks if not already available within their cloud domain. Databricks have natural support for MLFlow. This can be useful for timely transfer learning, model retraining and evaluation.

| MAJOR MILESTONES / DECISION POINTS: | Est. Hours | Est. Complete | Est. Act. Complete |
|---|---|---|---|
| Initial phases of discussion by BPC INT'L and Deloitte ABC_&_COs Chief Officers, CTOs, Project Consultants, etc. | This could vary depending on relationship between BPC INT'L and ABC_&_Co | This could vary depending on relationship between BPC INT'L and ABC_&_Co | This could vary depending on relationship between BPC INT'L and ABC_&_Co |
| In-house discussion between Deloitte ABC_&_COs Chiefs, CTOs and the Project Consultants. | Same as above | Same as above | Same as above |
| Initial ETL/ELT activities. Data engineering, data cleaning, data augmentation. Unity Catalogue arrangement if BPC INT'L cloud platform have support for unity catalogue. | 60 | 10/14/2024 | 10/25/2024 |
| ChemBERTa model fine-tuning through transfer learning.<br><br>Training with more BPC INT'L data set.<br><br>Initial model evaluation.<br><br>Discussion with BPC INT'L domain experts.<br><br>Further retraining and model fine-tuning based on domain experts' suggestions. | 80 | 10/25/2024 | 11/11/2024 |

# Implementation of ChemBERTa for Chemical Fingerprinting of Molecules and Property Prediction at BPC INT'L

**(Work Plan & Project Initiation (P.I.) Sheet)**

| | | | |
|---|---|---|---|
| Software and DevOp Engineers dockerization, cloud integration, Kubernetes deployment and possibly Desktop app design | 55 | 11/11/2024 | 11/19/2024 |
| Usage manual design by the software and NLP engineers<br><br>Training of BPC INT'L molecular engineers with regards to model/app usage. | 50 | 11/19/2024 | 11/26/2024 |
| Feedback from customers. ML and app refinements. | TBD | | |

**Project Time Line (Gantt Chart)**

| | **Approximately 7 weeks** | | | |
|---|---|---|---|---|
| | Week 1 (Approximately 1 week based on relationship between BPC INT'L and ABC_&_Co ) | **Approximately 6 weeks** | | |
| **Project Initiation** | Discussions/Project specification (BPC INT'L/ ABC_&_CO Team) | | | |
| **Data ETL/ELT activities** | | Data Engineers activities | | |
| **ChemBERTa Model transfer learning with BPC INT'L data** | | | NLP/ML engineers' activities | |

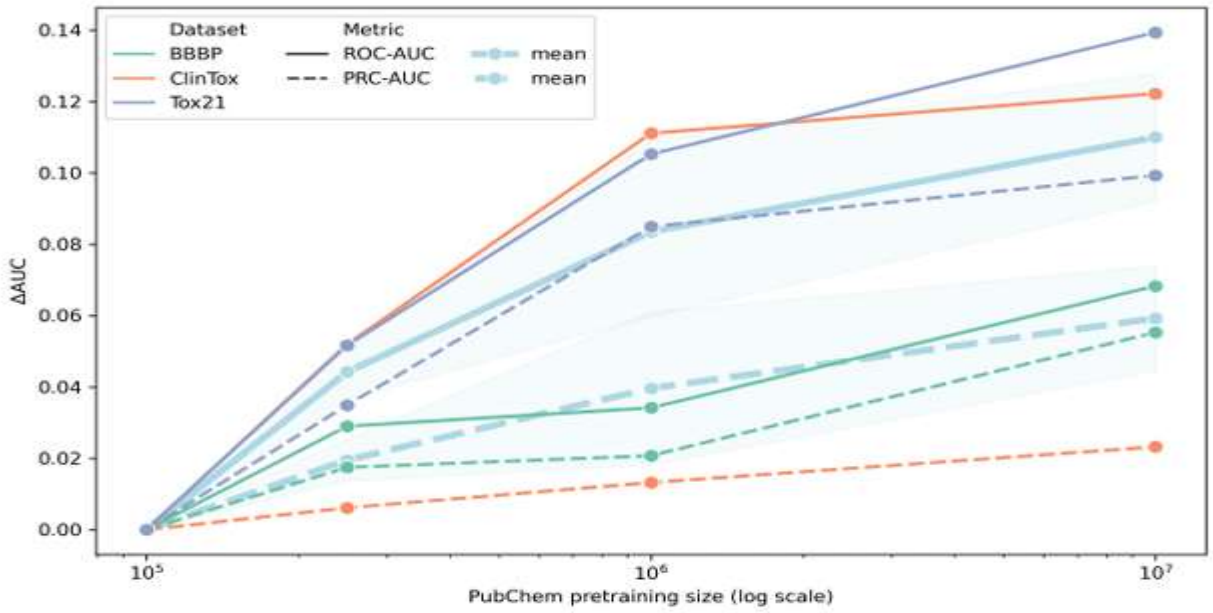
# Implementation of ChemBERTa for Chemical Fingerprinting of Molecules and Property Prediction at BPC INT'L

**(Work Plan & Project Initiation (P.I.) Sheet)**

| | | | | |
|---|---|---|---|---|
| **Software & DevOps Engineers** | | | | Model Deployment |

| **SUCCESS CRITERIA:** | |
|---|---|

- Model Evaluation: ROC and AUC curves as suggested by initial authors of ChemBERTa will be useful to measure developed models success. An example from the original ChemBERTa paper is shown below:




Model: Testing further with known molecules and their signature to see if model generalizes well. Apart from ROC/AUC curves and model generalizations, success criteria can also be based on the percentage of accurate molecular property predictions.

Success can also be measured by field reports from model users as BPC INT'L.

# Implementation of ChemBERTa for Chemical Fingerprinting of Molecules and Property Prediction at BPC INT'L

**(Work Plan & Project Initiation (P.I.) Sheet)**

| **PROJECT START UP APPROVAL SIGNATURES** | | | |
|---|---|---|---|
| Agreement on PI sheet content | | | |
| Project Consultant: | Emmanuel Oyekanlu, (ABC_&_CO) | Date: | |
| | | | |
| Key Customers: | BPC INT'L | Date: | |
| | | | |
| Project Sponsor: | BPC INT'L | Date: | |
| | | Date: | |