MolData, A Molecular Benchmark for Disease and Target Based Machine Learning

Arash Keshavarzi Arshadi (arashka@knights.ucf.edu)

Research article

Keywords: Artificial Intelligence, Benchmark, Biological Assays, Big Data, Database, Drug Discovery, Machine Learning, PubChem

Posted Date: October 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-968557/v1

Abstract

Deep learning's automatic feature extraction has been a revolutionary addition to computational drug discovery, infusing both the capabilities of learning abstract features and discovering complex molecular patterns via learning from molecular data. Since biological and chemical knowledge are necessary for overcoming the challenges of data curation, balancing, training, and evaluation, it is important for databases to contain meaningful information regarding the exact target and disease of each bioassay. The existing depositories such as PubChem or ChemBL offer the screening data of millions of molecules against a variety of cells and targets, however, their bioassays contain complex biological information which can hinder their usage by the machine learning community. In this work, a comprehensive disease and target-based dataset is collected from PubChem in order to facilitate and accelerate molecular machine learning for better drug discovery. MolData is one the largest efforts to date for democratizing the molecular machine learning, with roughly 170 million drug screening results from 1.4 million unique molecules assigned to specific diseases and targets. It also provides 30 unique categories of targets and diseases. Correlation analysis of the MolData bioassays unveil valuable information for drug repurposing for multiple diseases including cancer, metabolic disorders, and infectious diseases. Finally, we provide a benchmark of more than 30 models trained on each category using multitask learning. MolData aims to pave the way for computational drug discovery and accelerate the advancement of molecular artificial intelligence in a practical manner. The MolData benchmark data is available at https://github.com/Transilico/MolData as well as within the supplementary materials.

Introduction:

In the last decade, Artificial Intelligence (AI) has played a major role in modern computational aided drug discovery (CADD). Major improvements in both structure-based and ligand-based virtual screening have been recorded by training smart systems capable of identifying hidden molecular patterns. Learning models for Ligand-Based Drug Discovery (LBDD), or non-structural drug discovery, have been truly revolutionary in multiple aspects of the early drug discovery process. Deep Learning (DL) models have demonstrated the ability to discover abstract features of small molecules, allowing for better screening of both cell-based and target-based CADD. Using conventional methods, scientists would need to screen every molecule in a library on a specific target or cell, which is expensive, labor intensive, and time consuming. Virtual screening algorithms have introduced more affordable and faster alternatives that eliminate most of the early drug discovery costs. However, despite advances in CADD, the accuracy of traditional molecular modeling methods in most cases had not been satisfactory, prior to the introduction of Machine Learning (ML). Automatic feature extraction from molecules, and learning of hidden features in a large molecular library, are just some examples of what AI has changed forever in the drug discovery field (1)(2).

One of the most important factors of a reliable model is its training data, and deep learning models utilize this data to automate both pattern extraction and the prediction of bioactive molecules (3)(4). In general, datasets that are large, more diverse, and less biased result in training smarter systems with better inner features, performance, and generalization. Therefore, the first goal for machine learning scientists should be identifying and curating the right dataset per disease state. In addition, understanding the biological knowledge behind a dataset is as important as the data quality. Since data curation, model training, and model evaluation are time consuming and tedious, it is crucial to know the exact applications of the biological target for the disease of interest. Current input datasets need to be improved upon. Firstly, biomedical datasets tend to be very biased and imbalanced based on the biological assay and the chemical library (5). Secondly, understanding the exact cellular and molecular mechanism of the assays requires expert knowledge that ML scientists or cheminformaticians might not possess. Without knowing the biological background of the data, it would be difficult to devise solutions for data balancing and model evaluation. This knowledge is also necessary for finding appropriate public datasets due to their complicated descriptions and goals. Lastly, the chemical diversity, druggability, and toxicity of the predicted molecules need to be investigated (6)(7). With the

emergence of AI in non-structural drug discovery, there has been a renewed need for cleaned and clustered public molecular databases with simple and sufficient biological information, including the proper disease and targets involved in each bioassay.

There are multiple molecular depositories containing millions of molecules and hundreds of thousands of bioassays for specific biomedical aims. PubChem bioassays, ChemBL datasets, and ChemSpider are among some of the most comprehensive and well-known examples (8)(9)(10). These databases collect large sets of molecular activity outcomes for specific cells or protein targets. Even though these databases are excellent resources for model training, curating, and discovering the right bioassays, categorizing assays from them based on disease, targets, and signaling pathways can often be challenging and non-intuitive. Therefore, the scientific community has been benchmarking datasets and methods with these depositories, and in-house databases, in order to facilitate their usage and accelerate the advancement of molecular machine learning. Researchers often curate, analyze, and publish datasets with intended targets for discovering specific patterns in bioactive molecules. One of the first examples would be the 'Merck Molecular Activity Challenge' which had 15 biological assay tasks. In this dataset, targets are selected based on their cellular pathway relevance (11). In toxicity field, the Tox21 dataset from National Center for Advancing Translational Science (NCATS) containing 12 specific assays for nuclear receptor (NR) and stress response (SR) signaling pathway has been one of the most popular sources for advancing different learning methods such as transfer learning, multitask learning, few-shot learning etc. (7)(12)(13). Additionally, the PCBA dataset (14) from MoleculeNet and Massively Multitask Learning projects provided more than 120 PubChem bioassays with diverse sets of targets. It consists of curated public datasets, metrics for evaluations, and an open-source library in python called DeepChem (15)(16).

Even though these benchmarks have served to aid cheminformatician and ML scientists in discovering candidate drugs and allowed for better modeling, their bioassays lack the essential information like disease and target relevance. In addition, they generally do not cover a diverse set of diseases with a high number of screening molecules. We believe a reliable and practical ML model can be designed based on a known set of targets for a specific disease, fulfilling the need for a benchmark dataset that provides a comprehensive set of related assays. MolData is one of the most comprehensive disease and target-based benchmarks for democratizing molecular machine learning. It consists of 600 diverse bioassays from PubChem which are curated and clustered into 16 different diseases and 14 unique protein target classes. More than 1.4 million distinct molecules are presented in this benchmark, which consists of more than 170 million molecular screening data points. MolData aims to assist in the discovery of better and more diverse candidate drugs via the meaningful aggregation of large datasets. In doing so, it can be one of the main sources for the ML and Data Science community to develop practical molecular machine learning models. To demonstrate the application of MolData, we have run a correlation analysis to investigate drug repurposing, from which we have discovered three sets of bioassays highly correlated in both active and inactive molecules. Lastly, we trained more than 30 different multitask learning-based models, each for a specific disease or target, and one for all bioassays combined. These models can serve as a baseline for the data science community in order to advance molecular machine learning and enable better drug discovery.

Results:

1 – Benchmark Creation Pipeline

The overview of the benchmark creation pipeline is depicted in Figure 1. The process started by downloading the descriptions and summary of each data-source from PubChem. Due to the large number of selected bioassays, computational methods were implemented to aid in the creation of the benchmark, and serve as guideline for the manual tagging of each bioassay. The assay descriptions were first grouped into 10 clusters using BioBERT (17)(18), and tagged using a similar disease entity recognition model. Having done so, each description was tagged manually with the

assistance of the computational model results. By tagging assays in clusters separately, the similar keywords used for tagging were easier to detect. Manual tagging resulted in sixteen different disease-based categories of data. In addition, we used chEMBL repository (10) to identify each task's target class. After assigning each bioassay to one or more disease and target categories, the benchmark was analyzed with multiple approaches. After assigning each bioassay to one or more disease categories using specific keywords, the benchmark was analyzed with multiple approaches, such as mapping the molecular domain. For the application of drug repurposing, we ran a correlation analysis on the data and discovered three sets of correlated bioassays. Finally, different multitask graph convolutional neural networks (GCNNs) (19) were trained in order to create a baseline for the performance of multitask learning models in each disease related category.

2 – Data Aggregation Results

MolData benchmark originates from 9 of the largest data sources on PubChem in terms of number of screened molecules and number of active bioassays (20), as shown in Table 1. Initially, these collected data contained more than 1,000 bioassays, which were then triaged to 600 bioassays (AIDs) after filtering datasets smaller than 100,000 molecules or 15 active molecules. We included the updated Tox21 source (21) with more than 55 different bioassays due to their applicability to drug screening. As seen in Table 1, the activity percentage of each screening task was usually less than 1%, showing the imbalanced nature of the screening datasets.

Table 1
Data source summary.

PubChem Source	Aid count	Active data points	Total data points	Activity percentage (%)	Unique active molecules	Total unique molecules	Unique molecule activity percentage (%)
Broad Institute	67	125627	22.2m	0.56%	85579	472858	18.1%
Burnham Center for Chemical Genomics	67	139021	21.9m	0.63%	77159	381794	20.21%
Emory University Molecular Libraries Screening Center	12	24195	2.47m	0.98%	20964	348231	6.02%
ICCB-Longwood Screening Facility, Harvard Medical School	11	8358	2.1m	0.39%	6656	564021	1.18%
Johns Hopkins Ion Channel Center	22	48545	6.8m	0.71%	35487	344497	10.30%
NMMLSC	42	48186	11.5m	0.42%	37949	369431	10.27%
National Center for Advancing Translational Sciences (NCATS)	174	720319	53.4m	1.35%	240096	592616	40.51%
The Scripps Research Institute Molecular Screening Center	148	275224	47.6m	0.58%	142055	920418	15.43%
Tox21	57	21475	0.47m	5.67%	4183	8743	47.84%

3- Data Description Domain

To better understand the diversity within the 600 gathered bioassays, the description of each assay was fed to a BioBERT model (17). This model, which is trained on a large corpus of biomedical text, can create meaningful representations from the description of each assay and map the domain which these descriptions cover. Figure 2 depicts this map after clustering, showing how descriptions from different sources can have similar context to each other (e.g., bioassays from John Hopkins Ion Channel Center and The Scripps Research Institute Molecular Screening Center in cluster 2) or be distinct from the rest (e.g., bioassays from Tox21 in cluster 9).

The same model trained on disease entity recognition was also used to identify disease related key words in each description (22). While each cluster had some degree of similarity in terms of the diseases covered within each domain, it was far from perfect in correctly dividing the data domain based on their disease categories. Therefore, manual tagging was performed using the clusters and the disease entities as guidance. This process included highlighting disease related words within each bioassay's description and using them as tags to represent each bioassay. The dataset descriptions as well as their highlighted words are available in supplementary material 1.

4- MolData

4.1. Data Summary

After collecting all the specific disease identifiers or key words, we clustered them into 16 different categories. These categories were selected after carefully investigating all disease related words and their counts. The categories are: 1) Cancer, 2) Aging, 3) Bacterial, 4) Viral, 5) Fungal, 6) Parasitic, 7) Cardiovascular, 8) Immunological, 9) Nervous System, 10) Diabetes, 11) Epigenetic and Genetics, 12) Pulmonary, 13) Obesity, 14) Metabolic Disorder, 15) General Infection, and 16) Toxicity. The count of assays for each disease category is shown in Table 2. Overall, MolData consists of 600 bioassays with 1.4 million unique molecules, with nearly half of the molecules possessing activity in at least one bioassay. Moreover, MolData contains 224 tasks belonging to 2 or more disease categories. The MolData benchmark data is available at https://github.com/Transilico/MolData. All molecules, binary labels and splits are available in one file (supplementary 2), with two mapping files containing the mapping of each bioassay to each disease category (supplementary 3) and to each target category (supplementary 4).

Table 2
Disease-based information for the MolData Benchmark

Tag	Aid Count	Active Data Points	Total Data Points	Activity Percentage (%)	Unique Active Molecules	Total Unique Molecules	Unique Molecule Activity Percentage (%)
All Diseases	600	1410950	168345532	0.84	672935	1429989	47.06
Cancer	236	575454	68649771	0.84	230049	1323311	17.38
Nervous System	174	378812	54753975	0.69	170353	651249	26.16
Immune system	129	322362	38418661	0.84	157333	579658	27.14
Cardiovascular	94	212162	28660627	0.74	124270	542902	22.89
Toxicity	54	48653	2452656	1.98	30936	487219	6.35
Obesity	53	90837	14516199	0.63	65993	545513	12.1
Virus	47	113946	14679312	0.78	81702	621945	13.14
Diabetes	43	61408	11645151	0.53	47830	543600	8.8
Metabolic Disorders	42	126772	9985491	1.27	70665	527382	13.4
Bacteria	40	132593	12314737	1.08	89554	1290782	6.94
Parasite	24	98950	7302206	1.36	75027	500228	15
Epigenetics, Genetics	23	92837	6815597	1.36	65244	439537	14.84
Pulmonary	19	45940	6122297	0.75	36467	524167	6.96
Infection	11	93444	3312920	2.82	63782	521473	12.23
Aging	10	9030	3079580	0.29	8527	511471	1.67
Fungal	7	9253	2147751	0.43	8824	444373	1.99

The composition of each data category is depicted in Figure 3; showing how combining data from each data source resulted in the creation of each category. This combination demonstrates one of the main motivations for this work's data aggregation, as each disease category has related bioassays with multiple data sources. Furthermore, some categories such as Aging and Pulmonary are unexplored compared to those like Cancer and Nervous System, when large screening data is examined. These categories were selected based on their importance and the number of occurrences.

The protein targets of MolData in Figure 4 were classified by either 1) direct mapping to the CHEMBL database, 2) finding highly similar target in CHEMBL, or 3) manual curation (See methods). From the 419 total unique targets in MolData, 296 were classified into 14 classes (Figure 4). Enzymes (167/296) (Enzyme (other) + Hydrolase + Protease + Kinase + Transferase + oxidoreductase + NTPase + phosphatase) are the most prevalent class, followed by membrane receptors (44/296) and nuclear receptors (25/296). The occupancy of target classes is also reflected in the total assays for each class. For example, enzymes constitute the most prevalent class among the targeted assays (182/383), followed by membrane receptors (85/383) and nuclear receptors (53/383). The assays are overall enriched in the "privileged" targets, that is, membrane receptors, kinases, nuclear receptors, and ion channels. These four classes have been historically the

most prevalent among approved drug targets (23), accounting for 70% of the total approved drugs. In our dataset, however, 199 assays (52% total) represent targets from classes other than membrane receptors, kinases, nuclear receptors, and ion channels. When counting the total unique targets, these historically “unprivileged” targets even give a higher representation of the dataset with 190 counts (64% total). Therefore, MolData captures a higher diversity in the target classes compared to those of the approved drugs.

There are, additionally, classes that are overrepresented by our dataset compared to the set of targets with available approved drugs. For example, NTPases are targeted by 76334 unique compounds (29% of the total compounds from targeted assays), while only 2% of drugs target NTPases. Additionally, epigenetic regulators represent the target of 51776 unique compounds (20% of the total compounds from targeted assays), while only 0.3% of drugs interact with this class of proteins (23). These higher hit rate in the targets of MolData compared to the approved drugs could imply the inherent low druggability of such target classes or the lower significance of the targets for pharmaceutical industries.

Table 3
Target-based information for the MolData Benchmark

Target	Aid count	Unique target count	Active data points	Total data points	Activity percentage (%)	Unique active molecules	Total unique molecules	Unique molecule activity percentage (%)
All Targets	383	296	862370	103440515	0.83	261715	675161	38.76
Membrane receptor	85	44	146956	25922533	0.56	91489	458818	19.94
Enzyme (other)	54	51	83657	16210090	0.51	57808	632142	9.14
Nuclear receptor	53	25	74776	6083509	1.22	42838	442487	9.68
Hydrolase	36	32	113185	10830324	1.05	66195	526391	12.57
Protease	29	26	37943	7965313	0.47	30619	606793	5.05
Transcription factor	27	18	53416	4775685	1.11	40067	503249	7.96
Kinase	24	23	38257	7369690	0.52	31327	377519	8.29
Epigenetic regulator	23	20	76793	6840095	1.12	51776	523904	9.88
Ion channel	22	14	37402	6745762	0.55	28853	511873	5.63
Transferase	18	17	43955	6279651	0.7	30432	519646	5.85
Oxidoreductase	10	8	33956	2953760	1.15	30054	432578	6.94
Transporter	9	8	15390	2538579	0.60	15046	369621	4.07
NTPase	6	5	114465	1981575	5.78	76334	439967	17.34
Phosphatase	5	5	8090	1693773	0.48	6913	368329	1.87

4.2. Molecular Domain

To investigate the diversity of the screened molecules, all collected molecules are represented as vectors using ECFP4 (24). The results after applying Principal Component Analysis (PCA) are shown in Figure 5. The color in this figure

represents the density at each point, with denser areas becoming darker. The resulting map shows that while the selected molecules can occupy a large area within the fingerprint domain, a large percentage of them reside within the dark area, denoting a large degree of similarity within most of the screened molecules.

4.3. Correlation Analysis, a Showcase for Drug Repurposing

Drug repurposing is the process of finding new applications for already approved molecular drugs. These new applications can be target or disease based depending on the specific case of study. For example, during an outbreak, drug repurposing could be the fastest and most efficient option due to a lack of information about the new virus/bacteria, while novel drug discovery and the drugs subsequent approval may take many years (25). In the case of SARS-COV-2, Remdesivir was discovered by Gilead Sciences via repurposing a very similar molecular analog designed for inhibiting the viral replication of Ebola, a deadly virus with an outbreak in 2014 (26). For this benchmark, we hypothesized that correlating bioassays screened on different sets of targets would provide interesting information for better and faster drug repurposing. Therefore, the correlation score between the molecule bioactivity labels were calculated using a Pearson correlation coefficient.

Between all categories, toxicity showed the highest correlation between tasks, which is understandable due the nature of toxicity and the close biological relationship between the assays. In Figure 6, correlation heatmaps are shown for Toxicity assays and all non-toxicity assays with a correlation of 0.5 or more which have different targets. The second chosen group indicates higher correlation can exist between the labels of bioassays from the same, or different sources. Two sets of correlating targets and a viral similarity were discovered through this analysis. The first set of targets with a high correlation were 1) NPC1 2) SMN1 3) ATAD5 4) Rab9 5) STAT1. NPC1 and Rab9, with a 98% correlation, are important players in cholesterol metabolism and Niemann Pick Disease Type C (supplementary material 5) (27)(28). AIDs 485297 and 485313 were designed to discover the activators of mentioned proteins using luciferase reporter assays. Their high correlation to assays targeting STAT1 or ATAD5, which are important in cancer and immune disorders (29)(30)(31), is a valuable finding by the analysis of MolData benchmark for drug repurposing. Another interesting discovery was infectious disease based, as molecules targeting the Lassa Virus and Marburg Virus showed a high correlation. The Lassa Virus is a single stranded RNA virus with a circular morphology from the family of Arenaviridae, and is cause of Lassa hemorrhagic fever (32)(33). On the other hand, the Marburg virus belongs to the family of Filoviridae, with a shepherd's crook morphology, and causes similar symptoms to the Ebola virus, with a fatality rate of ~50% (34)(35). Both bioassays used the viruses envelop glycoproteins on a pseudotype virus system. We were curious to see if there has been any candidate drug with promising potency against both viruses. Favipiravir is a pyrazine carboxamide derivative that has shown effectiveness against both the Lassa and Marburg viruses (36)(37). These data suggest that MolData would be valuable source for further drug repurposing investigations.

4.4. Benchmark Classification Modeling, a Showcase for Bioactivity Prediction

The data from each disease and target category, as well as the aggregation of all bioassays, are used as training inputs for GCNNs. The classification results are shown in Table 4 and Table 5 as the baseline for each category. These results are from the imbalanced (untransformed) test set, weighted to ignore missing data points for each task (weight of 0), then averaged across all tasks within each category. The detailed results for each model and bioassay are presented in supplementary material 6. These results show the baseline performance for multitask models, with ROC AUC serving as the most important comparison metric due to the imbalance nature of the data.

Table 4
Classification results on the test set of disease categories, averaged on all tasks within each category

Disease Benchmark	Accuracy score	Recall score	Precision score	ROC AUC score
All Tasks	72.62 %	67.17 %	4.45 %	0.7756
Cancer	73.31 %	63.91 %	3.62 %	0.7648
Nervous System	72.22 %	61.79 %	2.43 %	0.7389
Immune System	71.35 %	62.87 %	2.82 %	0.7532
Cardiovascular	67.88 %	63.69 %	2.22 %	0.7307
Toxicity	59.87 %	74.72 %	14.75 %	0.7324
Obesity	72.02 %	61.74 %	3.71 %	0.7406
Virus	73.99 %	59.89 %	2.57 %	0.7447
Diabetes	69.87 %	64.51 %	3.82 %	0.7412
Metabolic Disorders	70.95 %	59.81 %	5.28 %	0.7200
Bacteria	72.87 %	67.02 %	3.26 %	0.7764
Parasite	73.15 %	72.11 %	4.66 %	0.8046
Epigenetics-Genetics	74.30 %	56.84 %	3.75 %	0.6974
Pulmonary	58.38 %	68.09 %	2.14 %	0.6951
Infection	70.56 %	68.76 %	6.51 %	0.7679
Aging	80.10 %	55.94 %	1.38 %	0.7625
Fungal	79.69 %	51.95 %	1.90 %	0.7484

Table 5
Classification results on the test set of target categories, averaged on all tasks within each category

Target Benchmark	Accuracy Score	Recall Score	Precision Score	ROC AUC Score
All Tasks w/ Targets	71.9 %	67.68 %	4.44	0.7714
Membrane receptor	66.91 %	62.36 %	1.69	0.7051
Enzyme (other)	66.72 %	74.39 %	1.92	0.7871
Nuclear receptor	62.63 %	73.6 %	11.25	0.7483
Hydrolase	71.7 %	67.23 %	2.85	0.7774
Protease	72.41 %	67.33 %	2.47	0.7606
Transcription factor	71.59 %	66.59 %	10.07	0.7565
Kinase	65.41 %	56.9 %	1.39	0.6664
Epigenetic regulator	70.35 %	68.11 %	3.91	0.7865
Ion channel	67.58 %	59.85 %	1.76	0.7104
Transferase	79.51 %	66.39 %	3.36	0.8079
Oxidoreductase	78.28 %	65.11 %	4.49	0.7868
Transporter	67.66 %	48.54 %	1.67	0.6525
NTPase	82.09 %	42.3 %	19.03	0.7703
Phosphatase	73.24 %	68.72 %	1.91	0.796

There are 383 tasks within the overall dataset that have both a disease-related tag as well as a target-related tag. These tasks are used for training multiple models including models trained on each disease category, each target category, and aggregation of all tasks with or without targets. Due to the repetition in training, different models' performance on these shared tasks can be compared to assess which multitask learning model was able to perform the best on each task. The results from this comparison are shown in Figure 7.

As shown in Figure 7, combining all tasks results in a higher average ROC AUC with the model trained on all 600 tasks being the best performer for the majority of tasks. However, there are 159 tasks which had their best performing model trained on fewer tasks, such as the models trained on specific disease or target categories, or the model trained on all tasks with target tags. This demonstrates that multitask learning on fewer tasks may be beneficial in some scenarios.

Discussion:

One of the main topics worth discussing is bias within the dataset. MolData consists of roughly 170 million data points. However, this screening was performed on 1.4 million molecules, denoting that each molecule exists on average in nearly 117 assays. Since the data sources are different, this level of repetitiveness shows a large overlap of molecules within the original data sources. Furthermore, as seen from the results of the molecular domain mapping, many of the molecules lie within a small section of the fingerprint domain, emphasizing their similarity. Therefore, a degree of bias exists within the gathered dataset with similar molecules being screened for each assay and in all data sources. We speculate this bias is due to the traditional rules used for selecting molecules as candidates for screening. One effective way to increase the diversity of chemicals would be switching from screening synthetic libraries to natural-based ones.

Natural derived compounds have shown a higher hit rate with the potential of targeting unknown and complex biotargets (38).

Another important topic to consider is the benchmark modeling result. The model architecture was selected to be shared within all models; however, this is suboptimal, and hyper-parameter optimization can be performed to find better possible architectures for each data category. This can apply to other hyper-parameters such as learning rate and batch size, which can be improved via a grid-search hyper-parameter optimization. Lastly, the low precision of the models is a focus of improvement since precision plays an important role in selecting molecules for future screening at inference time, directly affecting the cost and time of screening.

Conclusion:

MolData is one of the largest efforts in the collection, curation, and categorization of labeled molecular datasets. It consists of roughly 170 million screens of 1.4 million unique molecules distributed in 600 different bioassays and 16 disease categories, from cancer to infectious diseases. It also consists of a state-of-the-art target benchmark with 14 categories. We explored all the disease and target-related details in each bioassay for the development of a comprehensive benchmark to assist data scientists and the ML community in improving model development and computational drug discovery. We believe a key feature of any learning system is the training data, and the validation of a model is only possible with appropriate molecular and biological knowledge of the dataset. MolData takes advantage of a greater amount of labeled data compared to other benchmark datasets, which is an important addition to CADD. It is beneficial for the data science community to have a similar dataset for comparison of model performances; therefore, baseline performance is presented for 32 different categories. MolData hopes to take a step in furthering the molecular machine learning revolution, by providing the means for drug discovery and model development.

Methods:

A - Data Aggregation

The dataset was collected from PubChem bioassays due to its comprehensiveness and the high diversity of diseases and targets. We started with the selection of PubChem sources with highest number of Live Bioassays Counts and screened molecules. Hence, we selected eight sources including the 1) National Center for Advancing Translational Sciences, 2) Broad Institute, 3) Sanford-Burnham Center for Chemical Genomics, 4) NMMLSC, 5) Emory University Molecular Libraries Screening Center, 6) Tox21, 7) The Scripps Research Institute Molecular Screening Center, and 8) Johns Hopkins Ion Channel Center. As the final goal of this article is providing the machine learning community with a large, clustered dataset, we decided to include bioassays containing 100,000 or more molecules screened, as well as bioassays with more than 15 unique active molecules. This threshold was not applied to the Tox21 assays, which have a lower number of screened molecules, which were selected due to the importance of toxicity prediction to drug discovery. Table 1 shows the exact number of each sources' count, as well as active/inactive molecules.

B – Mapping the Data Domain with Natural Language Processing

After the assays are gathered and filtered by a size threshold, the process of understanding the context of the assays begins. Each assay contains information including the title of the assay, a general description, and optionally the biological target of the screening. To understand the diversity of the assays and map the domains which they cover, the description of each assay is analyzed using natural language processing tools, as elaborated upon in the following subsections.

B.1. Description Pre-Processing

The description of each bioassay was acquired from the PubChem website. Each description can contain a complete molecular and biological background, goal of each assay, and finally a brief description of the biological assay. However, each description may also contain unusable information such as the affiliated center, references, scientists involved in the screening, and grant information. Manual rules were written for of each of the eight data sources to filter out the lines containing the unusable information, resulting in cleaned descriptions explaining the assays' goal.

B.1. Feature Extraction using BioBERT

The cleaned descriptions were then lower-cased and fed to a BioBERT model for feature extraction. This model is trained for language modeling on a plethora of biomedical literature and can generate meaningful representation from biomedical text. Leveraging this capability, each description is transformed to a numerical vector of size 2048, representing what each assay's description contained. One disadvantage of this technique is the limited input size of BioBERT (512 token), which resulted in concatenation of some of the descriptions.

B.2. Clustering

Having acquired feature vectors of assay descriptions, they are clustered using K-Means clustering. Since the target of this clustering is to explore the domain which the descriptions cover, the number of clusters are unknown. To find the optimum number of clusters, the sum of squared distances of data points to their closest cluster center (SSE) are calculated and plotted based on the number of clusters. The optimum number of clusters is then found by detecting the knee point of the plot.

C – Tagging the Assays

After distinct clusters are formed from assay descriptions and the domains covered by the datasets are better defined, different assays can be grouped together to form a benchmark. The main form of distinction between the assays chosen in this work is disease relations. As previously mentioned, it is important for a dataset to provide each bioassay with simple disease and target labels for better computational drug discovery. To find the related diseases for each assay, the process of tagging is used, during which certain words in the description are chosen as tags to represent the assay. This process was implemented both using AI assistance and manual annotation.

C.1. BioBERT Disease Entity Recognition

The first approach implemented in this work to extract the disease related words from the description text of an assay is using a BioBERT model trained for disease entity recognition. This model takes a text sequence as input and returns the entity class related to each token, with the classes consisting of disease and non-disease entities. Using this model, all related disease keywords are extracted from each assay, automating the process of tagging. However, one major disadvantage of this technique is that many words within the description are disease related, but not defining for that assay. As an example, a task would claim that an older drug for a specific virus would be a carcinogen, falsely adding a disease tag related to "cancer" to the assay. The mentioned assay would have nothing to do with cancer, and was just an effort for antiviral drug discovery.

C.2. Manual Tagging

Since many of descriptions contain some biomedical-related words that are not defined for that specific task, understanding the exact biological assay and diseases related to the screening are crucial for tagging. A bioassay description contains a large amount of information regarding the target, related diseases, other proteins/RNAs/DNA down or up-stream, and in some cases the experimental details of the bioassay. In Figure 4, we provide a description from BioBERT cluster zero for AID 1259313 from Burnham Center for Chemical Genomics entitled "uHTS identification of small molecule modulators of NR3A". As shown in this figure, we first read the description for better understanding the

assay as a whole, as well as the tags found by the computational method, and then highlight any words with the potential of directing us to a special disease category. Here, Central Nervous System (CNS), Down Syndrome, and Neurological Disorders are the main words that direct us to the subcategories of 'Nervous System' and 'Epigenetics-Genetics'.

Activity of N-methyl-D-aspartate subtype of glutamate receptor (NMDAR) is essential for normal central nervous system (CNS) function. However, excessive activation of NMDAR mediates, at least in part, neuronal or synaptic damage in many neurological disorders, including hypoxic-ischemic brain injury and in Down syndrome. The dual role of NMDARs in normal and abnormal CNS function imposes important constraints on possible therapeutic strategies aimed at ameliorating or abating developmental disorders and neurological disease: blockade of excessive NMDAR activity must be achieved without interference with its normal function. We propose an approach for NMDAR modulation via modulation of the NR3A subunit, a representative of a novel family of NMDAR subunits with the goal to modulate the NMDAR activity. NR3 subunits have a unique structure in their M3 domain forming part of the channel region that contributes to decreased magnesium sensitivity and calcium permeability of NMDARs. It potently and specifically binds glycine and D-serine, but not glutamate. In addition, we have shown that glycine binding to the ligand-binding domain (LBD) of NR3A is essential for NR1/NR3 receptor activation, as opposed to internalization caused by ligand binding to NR1 LBD.

D – Benchmark Creation

After the disease related words are highlighted and extracted, each assay can be represented by its tags. The next step of the process is to use these tags for grouping related assays together, and to create the benchmark. To do so, major disease categories were first identified which could encompass all tags; and second, each tag was assigned to one or more related major disease. The relation between each tag and the major disease category can be found in supplementary material 7.

The classification of the protein targets of our dataset was gleaned by searching against the CHEMBL 29 (10) database. To classify the targets missing in CHEMBL, an all-by-all pairwise alignment was performed between MolBio targets and CHEMBL 29 dataset using phmmer 3.3 (39). If the top-scoring phmmer hit from CHEMBL aligns to the query sequence with a bit score of at least 100 and shares more than 80% similarity in sequence length, the classification is copied from the CHEMBL hit. The targets that neither mapped to CHEMBL or aligned confidently to CHEMBL using the mentioned criteria were annotated manually. The dataset originally contained 17 classes, but the list was curtailed to 13 (14 with the inclusion of "other") classes to remove the ones with assay occupancy of fewer than 5.

D.1. Molecular Data Pre-Processing

Having populated the disease categories, the molecular data for each assay was downloaded in the form of SMILES and their related bioactivity. The SMILES (Simplified molecular-input line-entry system) input for each molecule was canonicalized with isomeric information included. Duplicate or missing SMILES entries were then deleted. Regarding the bioactivity of the molecules, the existing labels in all assays are "Active", "Inactive", "Inconclusive", and "Unspecified". For the sake of consistency, molecules with inconclusive and unspecified labels were removed, and active and inactive molecules were respectively labeled as 1 and 0.

After the datasets are aggregated and preprocessed, Extended-Connectivity Fingerprints (ECFP4) are used to represent each molecule as a binary vector of 1024 length. This fingerprint represents existence or non-existence of certain sub-graphs within each molecule, which in turn makes it a suitable similarity metric between the molecules. To find the diversity within all the collected molecules, Principal Component Analysis is applied to the fingerprint vectors, projecting the fingerprints into a 2D space. To highlight the denser areas within this 2D map, gaussian kernel density estimation is used.

D.2. Correlation Analysis

To find correlating bioassay, the bioactivity labels of all molecules are taken as representing vectors of each bioassay. To begin, the shared labels between two bioassays that are non-missing are found. The Pearson correlation coefficient is calculated between these two vectors. This process is repeated for all bioassays within each disease category, as well as all the data. The resulting matrices are depicted in the result section. In order to find interesting correlations, the bioassays with a correlation coefficient higher than 0.5 or lower than -0.5 are selected. If the AID number of these bioassays are within 5 of each other (neighbors), they are dismissed, because in most cases they are very closely related screens. The remaining bioassays are further examined to check for any biological cause for this correlation.

D.3. Classification and Performance Benchmark

After the data is categorized based on their related diseases, the data is split into training, validation, and test sets, with 80, 10, 10 percent shares respectively. This splitting is done after finding the scaffold of each molecule, and molecules with shared scaffolds are put into same splits. Splitting based on the scaffolds creates more distinct splits, making the problem of classification harder and more like real-world scenarios where the inference set can often have a different distribution than the training set. Having split the data, some tasks may have no positive data points in the smaller splits, which creates a problem for calculating performance metrics, therefore, those tasks are identified, and one of their positive datapoints from the training set is randomly moved to the smaller split.

The molecules are featurized and converted into graphs with the chirality included in the features. To assist the process, the training split is balanced using weight transformers that affect how the loss is aggregated, amplifying the effect of positive samples during training. The training split is used to train a GCNN in a multitask manner for each category, including one model trained on 600 bioassays combined. The parameters for training and the related model are shown in Table 6.

Table 6
– Parameters of the training model.

Parameter	Value	Parameter	Value
Split	Specified	Dropout	0.1
Featurizer	GraphConv	Initial Learning Rate	0.0001
Epoch Number	10	Batch Size	128
Graph Conv. Layers	[512, 512, 512]	Dense Layer Size	1024

The evaluation metrics for the training of the models selected in this work are accuracy, recall, precision, and Area Under the Receiver Operator Curve (ROC AUC). While accuracy is a palpable metric of performance, it is not suitable for comparing models in imbalanced scenarios, where ROC AUC can correctly represent performance. Moreover, recall and precision are important in evaluating virtual screening models, since recall denotes how many of the valuable active molecules were correctly predicted, while precision demonstrates how well the trained model can do at inference time, selecting active molecules from a plethora of possible candidates for screening.

Declarations

Availability of data and materials:

All data are available in supplementary materials as well as a GitHub repository at: https://github.com/Transilico/MolData

Competing interests

We declare no conflict of interest

Funding

There is no funding for this project

Authors' contributions

Arash Keshavarzi Arshadi wrote the biological and chemical sections, collected the data, and manually labeled them. Milad Salem implemented the algorithms to clean and categorize the data, trained models and wrote the data science and analysis related sections. Arash Firouzbakht clustered the data to target related ones and wrote the target benchmark section. Jiann Shiun Yuan provided guidance and advised the project.

Acknowledgements

We would like to thank Hani Goodarzi for his pieces of advice and ideas for this project. We also thank Jennifer Collins and Julia Web for their contribution in improving the written sections.

References

1. Deng D, Chen X, Zhang R, Lei Z, Wang X, Zhou F. XGraphBoost: Extracting Graph Neural Network-Based Features for a Better Prediction of Molecular Properties. J Chem Inf Model [Internet]. 2021 Jun 28 [cited 2021 Oct 7];61(6):2697–705. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.0c01489
2. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS et al (2018) MoleculeNet: A benchmark for molecular machine learning. Chem Sci 9(2):513–530
3. Minnich AJ, McLoughlin K, Tse M, Deng J, Weber A, Murad N et al. AMPL: A Data-Driven Modeling Pipeline for Drug Discovery. 2019 Nov 12 [cited 2020 Mar 31]; Available from: http://arxiv.org/abs/1911.05211
4. Duan Y, Edwards JS, Dwivedi YK (2019) Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. Int J Inf Manage [Internet]. Oct 1 [cited 2019 Jun 16];48:63–71. Available from: https://www.sciencedirect.com/science/article/pii/S0268401219300581
5. Hussin SK, Abdelmageid SM, Alkhalil A, Omar YM, Marie MI, Ramadan RA. Handling Imbalance Classification Virtual Screening Big Data Using Machine Learning Algorithms. Complexity. 2021;2021
6. Karim A, Mishra A, Newton MAH, Sattar A (2019) Efficient Toxicity Prediction via Simple Features Using Shallow Neural Networks and Decision Trees. ACS Omega [Internet]. Jan 23 [cited 2021 Oct 7];4(1):1874–88. Available from: https://pubs.acs.org/doi/full/10.1021/acsomega.8b03173
7. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Front Environ Sci [Internet]. 2016 Feb 2 [cited 2019 Jun 25];3:80. Available from: http://journal.frontiersin.org/Article/10.3389/fenvs.2015.00080/abstract
8. PubChem [Internet]. [cited 2021 Oct 7]. Available from: https://pubchem.ncbi.nlm.nih.gov/
9. ChemSpider | Search and share chemistry [Internet]. [cited 2021 Oct 7]. Available from: http://www.chemspider.com/

10. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F et al (2015) ChEMBL web services: Streamlining access to drug discovery data and utilities. Nucleic Acids Res 43(W1):W612–W620
11. Merck Molecular Activity Challenge | Kaggle [Internet]. [cited 2021 Oct 7]. Available from: https://www.kaggle.com/c/MerckActivity
12. Richard AM, Huang R, Waidyanatha S, Shinn P, Collins BJ, Thillainadarajah I et al. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. Chem Res Toxicol [Internet]. 2020 Feb 15 [cited 2021 Oct 7];34(2):189–216. Available from: https://pubs.acs.org/doi/full/10.1021/acs.chemrestox.0c00264
13. Unterthiner T, Mayr A, Klambauer G, Hochreiter S. Toxicity Prediction using Deep Learning. 2015;3(February). Available from: http://arxiv.org/abs/1503.01445
14. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z et al. PubChem's BioAssay Database. Nucleic Acids Res [Internet]. 2012 Jan 1 [cited 2021 Oct 7];40(D1):D400–12. Available from: https://academic.oup.com/nar/article/40/D1/D400/2903189
15. Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V (2015) Massively Multitask Networks for Drug Discovery. Feb;
16. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci [Internet]. 2018 Jan 14 [cited 2019 Jun 17];9(2):513–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29629118
17. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH et al. Data and text mining BioBERT: a pre-trained biomedical language representation model for biomedical text mining. [cited 2021 Oct 7]; Available from: https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz682/5566506
18. Devlin J, Chang M-W, Lee K, Google KT, Language AI. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
19. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular Graph Convolutions: Moving Beyond Fingerprints. J Comput Aided Mol Des [Internet]. 2016 Mar 2 [cited 2020 Apr 26];30(8):595–608. Available from: http://arxiv.org/abs/1603.00856
20. Data Sources - PubChem [Internet]. [cited 2021 Oct 7]. Available from: https://pubchem.ncbi.nlm.nih.gov/sources/#sort=Live-BioAssay-Count
21. Tox21 - PubChem Data Source [Internet]. [cited 2021 Oct 7]. Available from: https://pubchem.ncbi.nlm.nih.gov/source/824
22. Li J, Sun Y, Johnson RJ, Sciaky D, Wei C-H, Leaman R et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database J Biol Databases Curation [Internet]. 2016 [cited 2021 Oct 7];2016:68. Available from: /pmc/articles/PMC4860626/
23. Santos R, Ursu O, Gaulton A, Patrícia Bento A, Donadi RS, Bologa CG et al. A comprehensive map of molecular drug targets. Nat Publ Gr [Internet]. 2017 [cited 2021 Oct 7]; Available from: www.nature.com/nrd
24. Rogers D, Hahn M. Extended-Connectivity Fingerprints (2010 May) J Chem Inf Model 50(5):742–754
25. Keshavarzi Arshadi A, Webb J, Salem M, Cruz E, Calad-Thomson S, Ghadirian N et al. Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Development. Front Artif Intell [Internet]. 2020 Aug 18 [cited 2020 Nov 19];3:65. Available from: www.frontiersin.org
26. Eastman RT, Roth JS, Brimacombe KR, Simeonov A, Shen M, Patnaik S et al. Remdesivir: A Review of Its Discovery and DevelopmentLeading to Emergency Use Authorization for Treatment of COVID-19. ACS Cent Sci [Internet]. 2020 May 27 [cited 2021 Oct 7];6(5):672. Available from: /pmc/articles/PMC7202249/
27. Lamri A, Pigeyre M, Garver WS, Meyre D. The Extending Spectrum of NPC1-Related Human Disorders: From Niemann–Pick C1 Disease to Obesity. Endocr Rev [Internet]. 2018 Apr 1 [cited 2021 Oct 7];39(2):192. Available from:

/pmc/articles/PMC5888214/

28. K N AC, K D, DK S, EL H, DL M et al. Protein transduction of Rab9 in Niemann-Pick C cells reduces cholesterol storage. FASEB J [Internet]. 2005 Sep [cited 2021 Oct 7];19(11):1558–60. Available from: https://pubmed.ncbi.nlm.nih.gov/15972801/
29. Giovannini S, Weller M-C, Hanzlíková H, Shiota T, Takeda S, Jiricny J (2020) ATAD5 deficiency alters DNA damage metabolism and sensitizes cells to PARP inhibition. Nucleic Acids Res [Internet]. May 21 [cited 2021 Oct 7];48(9):4928–39. Available from: https://academic.oup.com/nar/article/48/9/4928/5820885
30. Pensa S, Regis G, Boselli D, Novelli F, Poli V. STAT1 and STAT3 in Tumorigenesis: Two Sides of the Same Coin? 2013 [cited 2021 Oct 7]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK6568/
31. Chapgier A, Wynn RF, Jouanguy E, Filipe-Santos O, Zhang S, Feinberg J et al (2006) Human Complete Stat-1 Deficiency Is Associated with Defective Type I and II IFN Responses In Vitro but Immunity to Some Low Virulence Viruses In Vivo. J Immunol [Internet]. Apr 15 [cited 2021 Oct 7];176(8):5078–83. Available from: https://www.jimmunol.org/content/176/8/5078
32. Richmond JK, Baglole DJ (2003) Lassa fever: epidemiology, clinical features, and social consequences. BMJ Br Med J [Internet]. Nov 29 [cited 2021 Oct 7];327(7426):1271. Available from: /pmc/articles/PMC286250/
33. Lassa fever [Internet]. [cited 2021 Oct 7]. Available from: https://www.who.int/health-topics/lassa-fever#tab=tab_1
34. OG G, BE J, WJ MRV, GW V T, HE L. Drug targets in infections with Ebola and Marburg viruses. Infect Disord Drug Targets [Internet]. 2009 Oct 30 [cited 2021 Oct 7];9(2):191–200. Available from: https://pubmed.ncbi.nlm.nih.gov/19275706/
35. Marburg virus disease [Internet]. [cited 2021 Oct 7]. Available from: https://www.who.int/news-room/fact-sheets/detail/marburg-virus-disease
36. Rosenke K, Feldmann H, Westover JB, Hanley PW, Martellaro C, Feldmann F et al. Use of Favipiravir to Treat Lassa Virus Infection in Macaques - Volume 24, Number 9—September 2018 - Emerging Infectious Diseases journal - CDC. Emerg Infect Dis [Internet]. 2018 Sep 1 [cited 2021 Oct 7];24(9):1696–9. Available from: https://wwwnc.cdc.gov/eid/article/24/9/18-0233_article
37. SL B, TM B, SA JWKSW VT, L D, et al. Efficacy of favipiravir (T-705) in nonhuman primates infected with Ebola virus or Marburg virus. Antiviral Res [Internet]. 2018 Mar 1 [cited 2021 Oct 7];151:97–104. Available from: https://pubmed.ncbi.nlm.nih.gov/29289666/
38. Li R, Npr /, Wilson BAP, Thornburg CC, Henrich CJ, Grkovic T et al. Natural Product Reports Creating and screening natural product libraries. 2020;37:863–1032
39. HMMER [Internet]. [cited 2021 Oct 7]. Available from: http://hmmer.org/

Figures

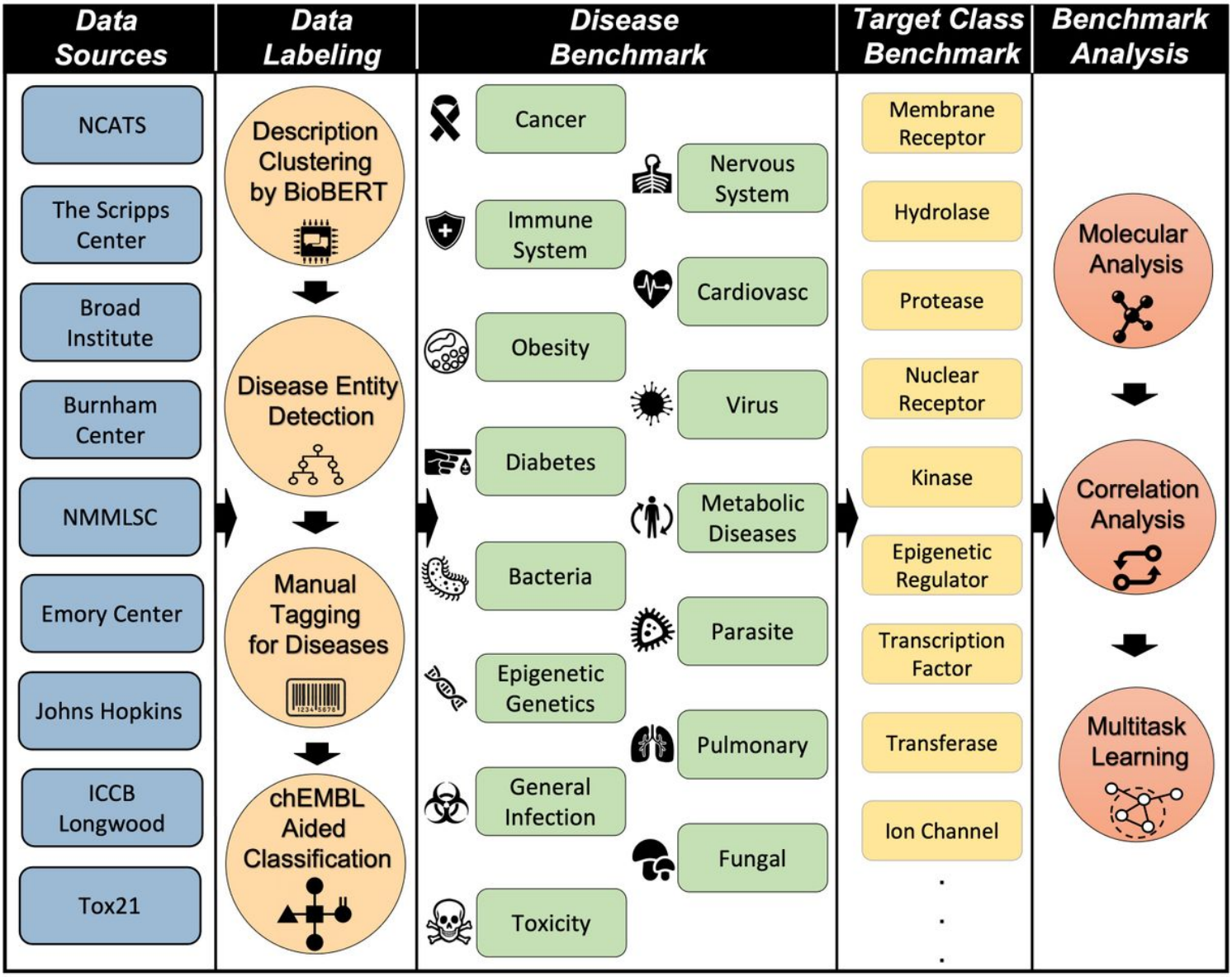

Figure 1

The pipeline of MolData benchmark creation.

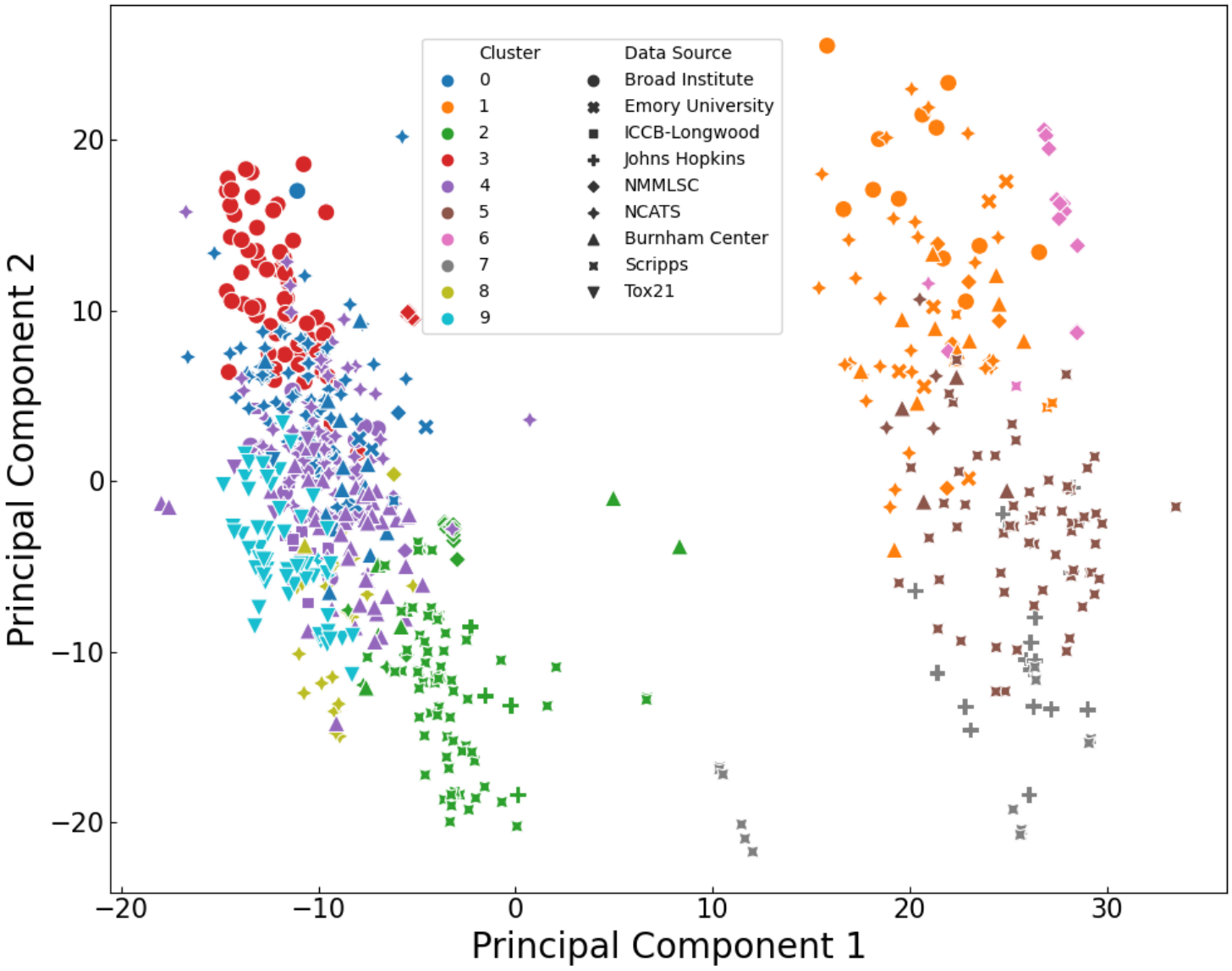

Figure 2

Map of the bioassays' descriptions using the output of the BioBERT model.

Figure 3

Number of bioassays for each disease category and their original source.

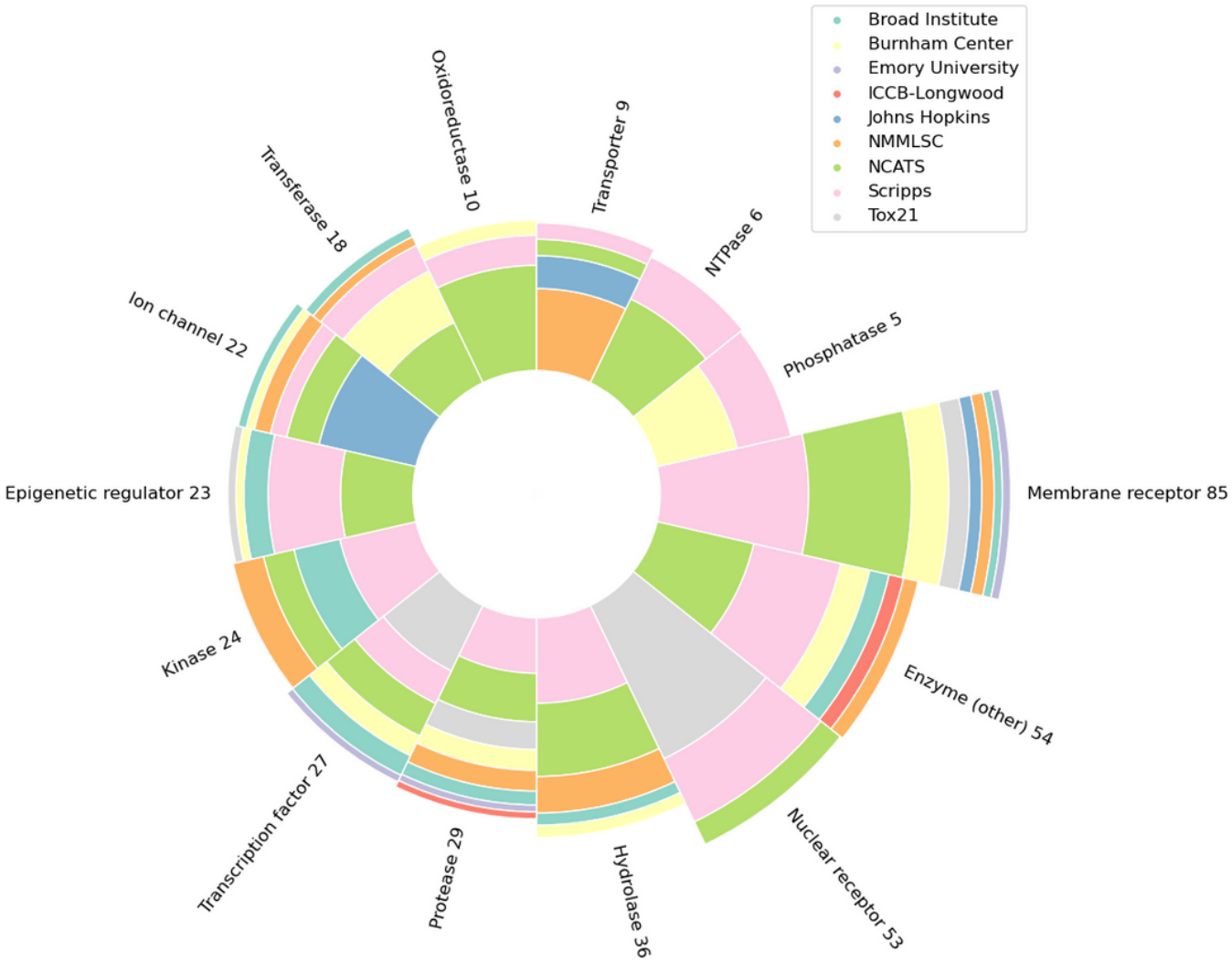

Figure 4

Number of bioassays for each target category and their original source.

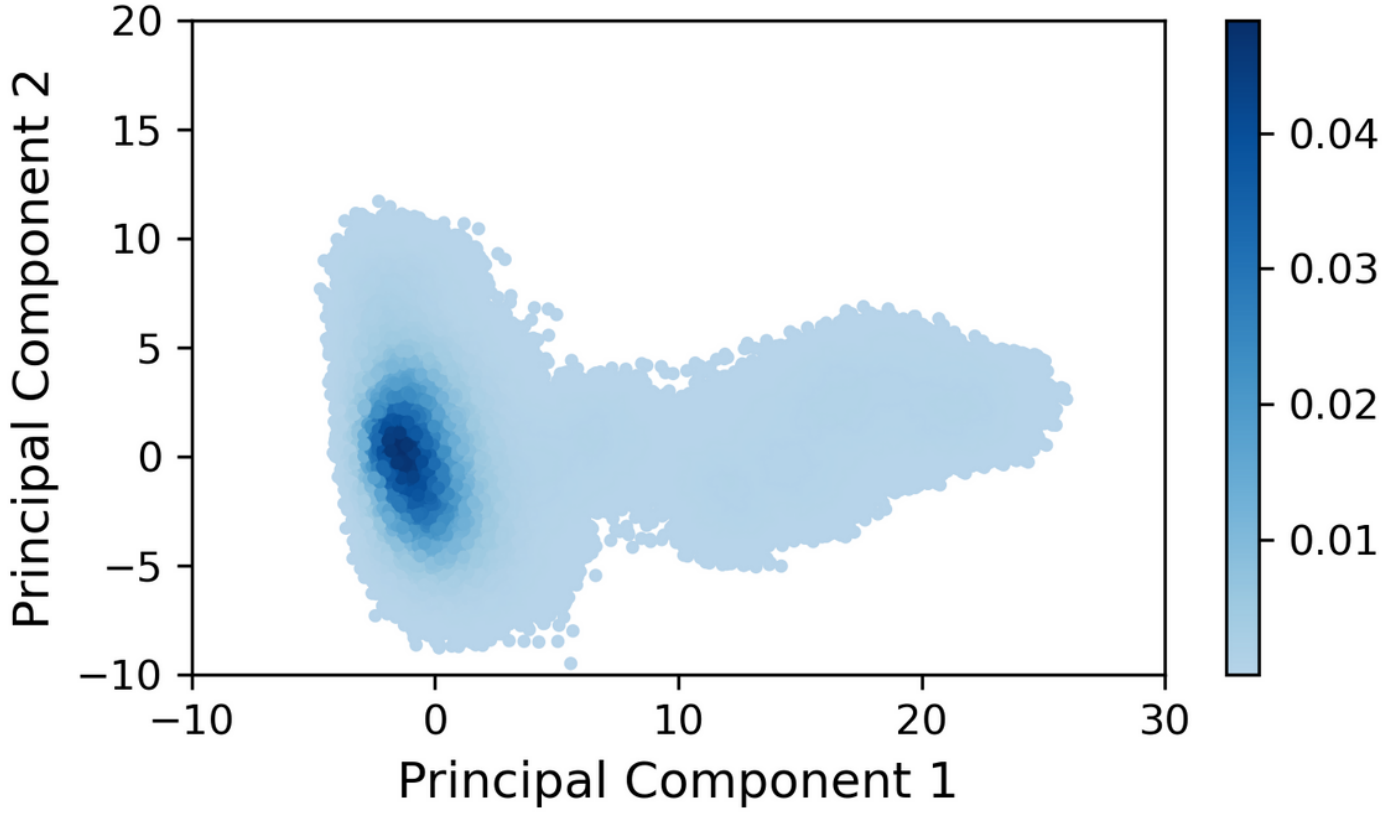

Figure 5

Gaussian kernel density estimation of the molecular fingerprint (ECFP4) space after projection to the two principal components.

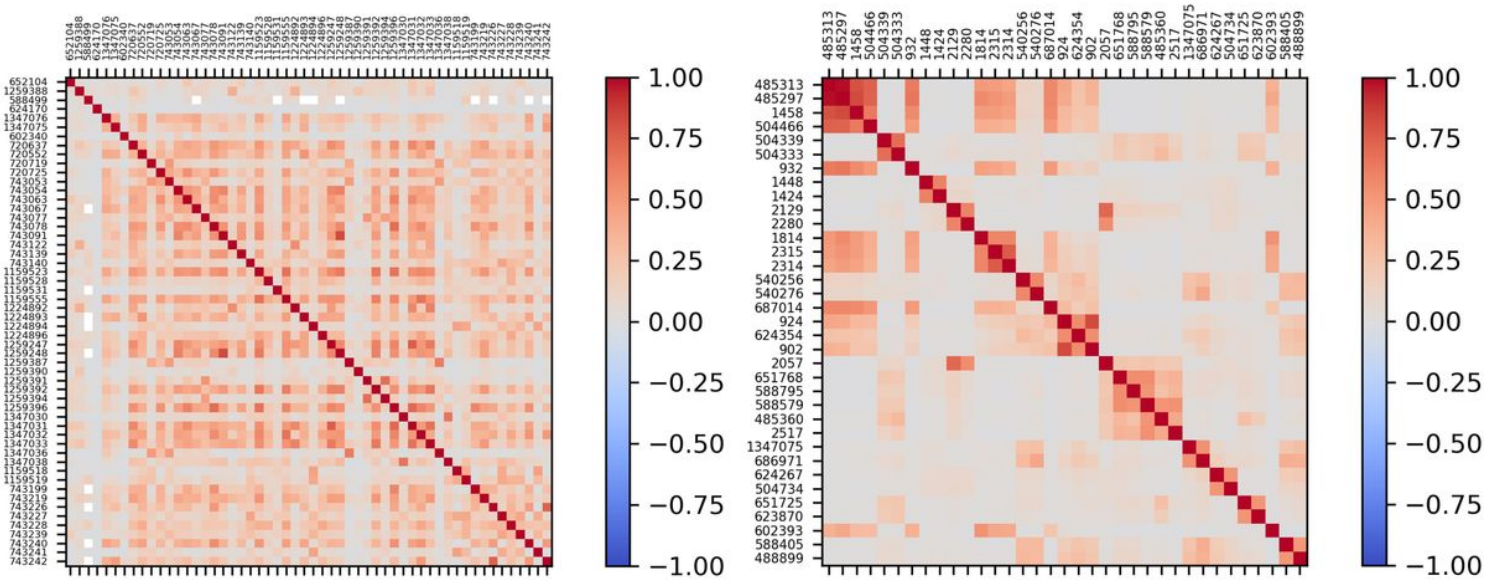

Figure 6

Correlation matrix between a) Toxicity bioassays, b) Non-Toxic bioassays with correlation > 0.5.

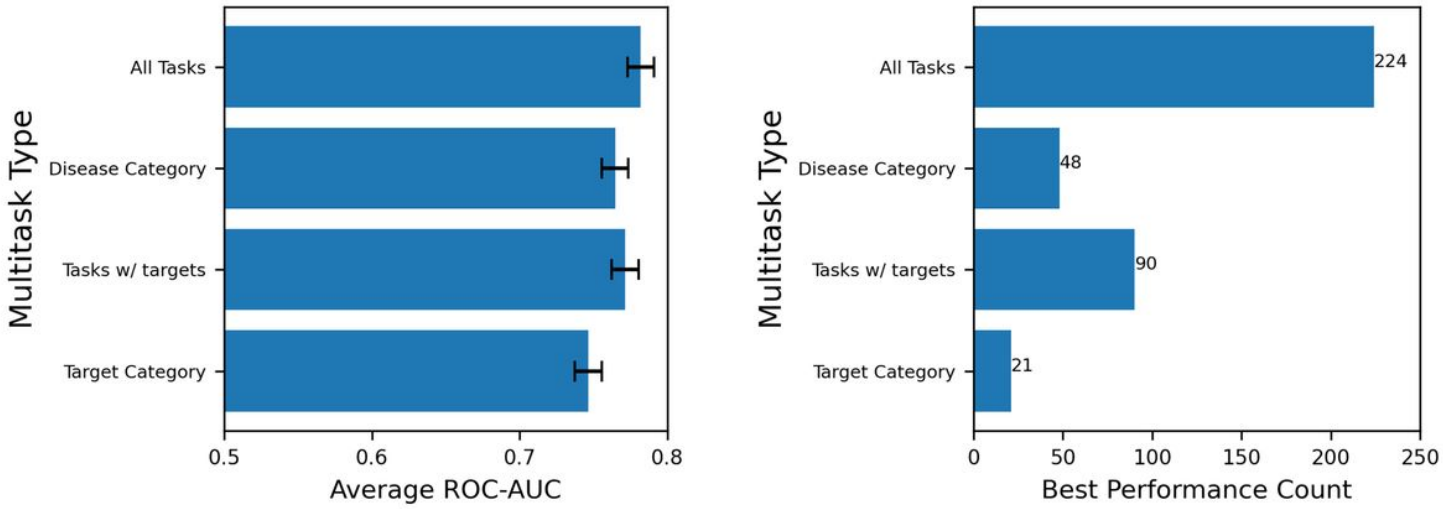

Figure 7

Comparison of different multitask models for 383 shared tasks in regard to a) average performance with 90% confidence interval, and b) number of tasks where each model was the best performing

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial1.csv
- SupplementaryMaterial2.zip
- SupplementaryMaterial3.csv
- SupplementaryMaterial4.csv
- SupplementaryMaterial5.csv
- SupplementaryMaterial6.xlsx
- SupplementaryMaterial7.csv