EnSidNet: Enhanced Hybrid Siamese-Deep Network for grouping clinical trials into drug-development pathways

Lucia Pagani
Analytics Center of Excellence, IQVIA
Via Fabio Filzi 29, Milan, Italy
lucia.pagani@iqvia.com

Abstract

Siamese Neural Networks have been widely used to perform similarity classification in multi-class settings. Their architecture can be used to group the clinical trials belonging to the same drug-development pathway along the several clinical trial phases. Here we present an approach for the unmet need of drug-development pathway reconstruction, based on an Enhanced hybrid Siamese-Deep Neural Network (EnSidNet). The proposed model demonstrates significant improvement above baselines in a 1-shot evaluation setting and in a classical similarity setting. EnSidNet can be an essential tool in a semi-supervised learning environment: by selecting clinical trials highly likely to belong to the same drug-development pathway it is possible to speed up the labelling process of human experts, allowing the check of a consistent volume of data, further used in the model's training dataset.

1 Introduction

Siamese Neural Networks (SNN) were developed in the early 1990s (Bromley et al., 1994) to obtain a similarity score from examples of signatures with the goal of identifying forgery. From then many applications used SNN, primarily on image recognition tasks (Chopra et al., 2005). The basic architecture of SNN consists of two identical networks able to learn the hidden representation of the inputs. A similarity function would then compare the inputs hidden representations. The similarity score was taken advantage of in contexts like 1-shot learning in multiclass-classification problems, where a single example of a class was seen by the algorithm only once before making inference (Koch et al., 2015).

Different architectures of SNN were developed in time: Simo-Serra and colleagues developed a 3-inputs SNN (Simo-Serra et al., 2015), where the neural network learned to rank the outputs and identify whether the reference's hidden representation is more similar to a positive or a negative sample.

Another example involves the insertion of an intermediate stage between the similarity score layer and the final prediction layer (Subramaniam, Chatterjee, and Mittal, 2016), allowing to increase performance in person re-identification task despite partial occlusion and difference in point of view or illumination.

The first applications of SNN were based on Convolutional Neural Networks (CNN) to obtain similarity score on images (Simo-Serra et al., 2015), seeing SNN involved in different tasks such as patch identification (Simo-Serra et al., 2015), person identification (Ahmed et al., 2015), image matching from different angles (Vo and Hays, 2016). SNN was also explored in Natural Language Processing (NLP) contexts in tasks like identifying sentence similarity (Mueller and Thyagarajan, 2016) and support relation for argumentation (Gema et al., 2017). These applications highlight the flexibility of SNN to identify similarities in different contexts. Here we apply this architecture on an unmet healthcare task: grouping clinical trials belonging to the same drug-development pathway.

Before being released on the market a new drug needs to go through several expensive and time-consuming experiments, involving testing the pharmacological characteristics of the drug in

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 254–266
June 6–11, 2021. ©2021 Association for Computational Linguistics

biochemical, cellular, and animal models (preclinical phase) and then on human volunteers (clinical stage). The clinical stage is divided into 3 pre-approval phases (safety, efficacy, regulatory proof) and a fourth post-market phase (Corr and Williams, 2009). The experiments performed by research or pharmaceutical companies to study a drug in human subjects are called clinical trials. A drug-development pathway is defined as all the clinical studies performed on a drug for an indication to obtain approval from the regulatory agency. Example of a drug-development pathway is presented in Supplementary Table 1. From starting a phase 1 clinical trial to obtaining approval from a regulatory agency, a drug can be tested for over 10 years, and the process can cost hundreds of millions of dollars, involving thousands of subjects, including patients, doctors, nurses and other personnel, with an approval rate of around 10% (Wong, Siah, and Lo, 2019).

Information on most clinical trials is publicly available. Pharmaceutical companies are asked to share their information on ClinicalTrials.gov, a U.S. National Library of Medicine resource. Other companies such as DrugBank (Wishart et al., 2006) or Citeline (Wong, Siah, and Lo, 2019) parse the information from ClinicalTrials.gov and add a hand-curation process in which human labellers cross-reference certain information and add additional labels to the trials, resulting in a similar but more accurate database.

Although having information on the clinical trials related to the development of a drug may seem a very straightforward process, there are many confounding factors:

- Very often several trials of the same phase are run, to obtain statistical power or on slightly different protocols (country, population, sample size, …)
- The same trial can belong to two different phases (e.g. phase 1-2 or 2-3)
- The company may not share on public databases the information of the trials it is performing, or may share partial information or not update them
- Some phases may be skipped
- Often subsequent trial phases from the same drug-development pathway may address slightly different diseases
- The disease and the drug can be referred to from different nomenclatures in different trials

Grouping of clinical trials to the same drug-development pathway is a requirement for many different applications, such as analyzing the success of a pharmaceutical company performing trials and marketing new drugs, or calculating the probability of success of a drug for a therapeutic area, evaluating the number of pathway in a therapeutic area, and investigating the futility of a pathway.

Although there is a strong need for a large freely-available dataset, only proprietary hand curated datasets exist (Wong, Siah, and Lo, 2019). A relatively small dataset of regulatory agency approved pivotal trials could be parsed from Food and Drug Administration Drug Trials Snapshots (FDA Snapshot) (https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots). The lack of large publicly available datasets may be one of the reasons why to our knowledge no algorithms to group clinical trials in drug-development pathways have been described in the literature.

The contributions of this paper are: (a) a novel approach to group clinical trials in drug-development pathways; (b) an iterative semi-supervised learning pipeline to optimize the grouping of clinical trials to the pathway.

The model proposed here is based on a SNN architecture. The model learned the similarity of trials belonging to the same pathway. The advantage of using the proposed model in a semi-supervised learning pipeline would lead to decreased human-labelling effort; the proposed pipeline can work in a *de-novo* mode (fresh start) and in a primed mode (adding data to previously scored pathways).

2 Methods

2.1 Data used to train and validate model

The ground truth pathways considered in this experiment were pathways extracted by the pivotal trials from the FDA Snapshot and manually identified pathways (hand-curated). For more details on the datasets composition and other methods considered here see Supplementary Methods.

2.2 Neural Network architectures

Three architectures were compared in the current research, schematized in Supplementary Figure 1: pure Siamese Neural Network architecture (SNN) where only Siamese branches were present, a hybrid Siamese and Deep Neural Network (SiD NN) consisting of Siamese character-based branches and an additional input branch, and an enhanced version of the SiD NN, having a fully connected layer before the prediction layer (EnSidNet). Supplementary Methods contain the detailed description of the 3 architectures.

2.3 Inputs of the model

The input features of the networks were: the drugs used in the clinical trial (intervention), the disease considered (condition), the phase of the trial (phase), the countries where the clinical trial was conducted (country), the sponsors of the trial (sponsor), the start and end date of the trial (expressed in days compared to an arbitrary reference date, January 1st 2000). Details of the preprocessing of the inputs can be found on Supplementary Methods.

2.4 Prediction Algorithm

Algorithm 1 contains the pipeline to apply the Neural Network to group trials into pathways.

Algorithm 1

```
Input: trials to group in pathways and previously scored pathways
Output: pathways containing development trials
1:  divide trials in therapeutic areas
2:  for every therapeutic area do
3:      for every existing pathway do
4:          predict similarity between 2 trials of a present pathway and a new trial
5:          if probability > 0.8 for both couples do
6:              add trial to present pathway
7:      sort trials (common lead sponsor or condition)
8:      divide trials into batches
9:      for every trial in batch do
10:         match all versus all and predict similarity
11:         if probability > 0.8 do
12:             group the trials in a pathway
13:     group pathways with common trial
14:     select 1 trial per pathway and repeat steps 9-13
15: return pathways
```

Algorithm 1

The details of the pipeline are reported in Supplementary Methods. For schematic example of the matching pipeline see Supplementary Figure 2.

3 Experiments

In Supplementary Table 2 we report the number of parameters of the networks and training time. The three neural models have different number of parameters to train, and the complexity of SNN compared to the hybrid models made the training time per epoch longer. In terms of time per epoch the other two hybrid models had comparable time per epoch, despite the slightly higher complexity of EnSidNet compared to SiD NN.

3.1 Balanced datasets

Accuracy was tested on a balanced validation dataset (see dataset splitting for details on balanced dataset creation). It can be seen from Table 1 that the best performing algorithm was EnSidNet.

	Balanced dataset
	Accuracy
SNN	0.763393
SiD NN	0.907738
EnSidNet	**0.91369**

Table 1: Accuracy of the best model on a balanced dataset

3.2 32-way 1-shot evaluation performances

One-shot evaluation was used to predict whether a new trial belongs to established pathways.

The score expected from a random classifier is 3.125, due to the unbalanced 1:32 ratio of positive couples versus negative. It can be seen in Table 2 that all neural models scored significantly higher than a random classifier in a 32-way 1-shot evaluation assay.

	32-way 1-shot evaluation assay		
	Neural Network	**1-Nearest Neighbor**	**Random Classifier**
SNN	66.67	81.82	6.06
SiD NN	93.94	69.70	0
EnSidNet	**96.97**	69.70	3.03

Table 2: Results of 1-shot evaluation assay

EnSidNet was the model with the highest performance in the test set. On the contrary, the SNN had the lowest performance between the neural models. Surprisingly the input format of SNN tested on the heuristic 1-Nearest Neighbor gave a relatively high performance.

To understand the contribution of the different features on the final EnSidNet prediction a SHAP analysis was performed. As Supplementary Figure 3 shows the most important feature to distinguish between couples from the same or different pathway is the number of common sponsors. It is interesting to note that the most contributing features belong to the additional inputs branch of the NN, features that increased the performance of the 32-way 1-shot learning metric of almost 30% (see Table 2).

3.3 Metrics on imbalanced dataset

Table 3 shows the other metrics considered in this research, calculated on the 1:32 unbalanced dataset.

	Unbalanced dataset				
	F1	P	R	ROC AUC	PR AUC
SNN	0.16	0.09	0.76	0.85	0.61
Sid NN	**0.90**	**0.86**	**0.94**	0.97	0.89
EnSidNet	**0.90**	**0.86**	**0.94**	**0.99**	**0.92**

Table 3: Metrics of the neural models. P = Precision, R = Recall, ROC AUC = area under Receiver Operating Curve, PR AUC = area under Precision-Recall curve

SNN had the worst performance on all metrics. Despite Sid NN had performances comparable to EnSidNet on precision and recall, ROC AUC and PR AUC showed the higher performance of the Enhanced model.

Figure 1 shows the probabilities associated to couples belonging or not to the same drug-development pathway for EnSidNet. The figure shows that the algorithm can distinguish with great certainty whether the trials belong to the same pathway or not, and the higher recall than precision.

3.4 Trials grouping in pathways

Algorithm 1 for grouping the trials in possible pathways was applied to clinical trials present in the DrugBank database. The clinical trials included were those in phases 1, 2 and 3, with industry lead sponsors and 'treatment' as the purpose of the trial. Trials to match into drug-development pathways were 34188. The algorithm took less than 4 hours to run.

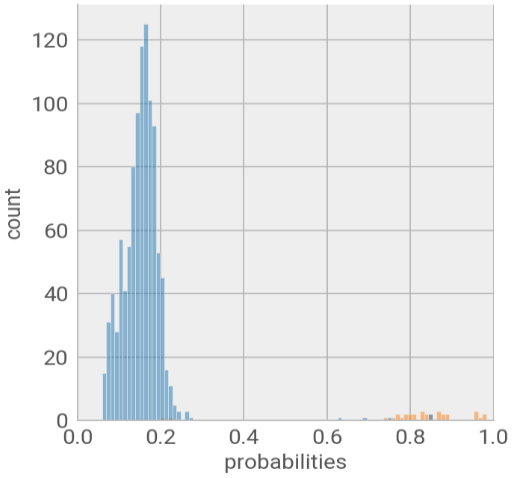


Figure 1: Predictions probability distribution. Blue bars represent couple of trials from different pathways, orange trials from the same pathway

The therapeutic areas included in these pathways were 27.

As presented in Table 4 the statistics of the possible pathways obtained from Algorithm 1 is overlapping with the statistics of the datasets used to train the neural networks (Supplementary Table 3).

	# pathways per therapeutic area	# trial per pathway
min	0	2
25 percentile	2.5	2
50 percentile	7	2
75 percentile	9.5	3
max	26	49
total	191	629 (583 unique)

Table 4: Statistics of the possible pathways obtained by running EnSidNet

Despite the input of Algorithm 1 was more than 34,000 trials, less than 600 were matched in pathways. However, the possible pathways obtained were about 1.5 times the number of total pathways in the dataset, suggesting new possible pathways were discovered running Algorithm 1, highlighting the potential of this semi-supervised approach for the grouping of clinical trials in pathways.

A subset of the predicted pathways was given to human labellers for scoring. The 73 predicted pathways (2-49 trials long), for a total of 264 trials, gave rise to 165 different trials (1-11 trials long). The different distribution of the predicted versus confirmed pathways can be seen in Supplementary Table 4. A total of 112 trials (42%) were confirmed being assigned by the algorithm to proper pathways. Only two of the trials selected for human scoring were found also on the ground truth datasets. Specifically, both trials belonged to the FDA snapshot dataset and were single-trial pathways. Interestingly, one of these trials was assigned to 2 other trials, and this 3-trial pathway was then confirmed by the human experts scoring. This is a good example of the capability of EnSidNet and the proposed algorithm to find the contributing trials to a drug-development pathway.

4 Conclusion

We present a new approach for the grouping of clinical trials into drug-development pathways. To meet this objective, we proposed 3 different neural network architectures. The best performing model was EnSidNet, an enhanced hybrid Siamese-Deep Neural Network.

EnSidNet was used to develop a semi-supervised learning pipeline using 1-shot evaluation and classification to group trials into existing or new pathways. Human scoring would lead to the increase of the training size with *ad-hoc* positive and negative samples.

Acknowledgements

We thank Gregory Lever and Joseph Heenan for useful feedback.

References

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, Roopak Shah. 1994. Signature verification using a "Siamese" time delay neural network. In *Advances in Neural Information Processing Systems, 6:737–744.*

Sumit Chopra, Raia Hadsell, Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1:539–546.*

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image Recognition. In *Proceedings of the 32 nd International Conference on Machine Learning, 37.*

Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, Francesc Moreno-Noguer. 2015. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*:118-126.

Ejaz Ahmed, Michael Jones, Tim K. Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*:3908-3916.

Nam Vo, James Hays. 2016. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision*:494-509.

Arulkumar Subramaniam, Moitreya Chatterjee, Anurag Mittal. 2016. Deep Neural Networks with Inexact Matching for Person ReIdentification. In *Advances in Neural Information Processing Systems*:pp. 2667-2675.

Jonas Mueller, Aditya Thyagarajan. 2016. Siamese Recurrent Architectures for Learning Sentence Similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*.

Aryo P Gema, Suhendro Winton, Theodorus David, Derwin Suhartono, Muhsin Shodiq, Wikaria Gazali. 2017. It Takes Two To Tango: Modification of Siamese Long Short Term Memory Network with Attention Mechanism in Recognizing Argumentative Relations in Persuasive Essay. In *Procedia Computer Science*, 116:449-459.

Peter B Corr, David A Williams. 2009. The Pathway from Idea to Regulatory Approval: Examples for Drug Development. In *U.S. National Library of Medicine.*

Chi H Wong, Kien W Siah, Andrew W Lo. 2019. Estimation of clinical trial success rates and related parameters. In *Biostatistics, 20(2):273–286.*

David S Wishart, Craig Knox, An C Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, Jennifer Woolsey. 2006. Drugbank: a comprehensive resource for in silico drug discovery and exploration. In *Nucleic Acids Research, 34:D668-72. 16381955.*

EnSidNet: Enhanced Hybrid Siamese-Deep Network for grouping clinical trials into drug-development pathways

Lucia Pagani
Analytics Center of Excellence, IQVIA
Via Fabio Filzi 29, Milan, Italy
lucia.pagani@iqvia.com

A. Supplementary Methods.

Training and evaluation of the models was run on a 4 CPU/32 GiB RAM machine, while Algorithm 1 was run on an 8 CPU/64 GiB RAM machine.

Pathway dataset

Supplementary Table 3 shows the statistic of the pathways in the two ground truth datasets: FDA Snapshot and hand curated.

Dataset splitting

A 4 folds split was performed in this research:

- Training and validation set: split in 80% for training and 20% for validation, it was composed of balanced couples of trials belonging and not belonging to the same pathway
- 32-way 1-shot evaluation validation set: this dataset was composed by 1 couple of trials belonging to the same pathway and 31 randomly coupled trials
- 32-way 1-shot evaluation test set: similar to the previous dataset, this dataset contained only 1 couple belonging to the same pathway over 32 randomly chosen couples of trials

Supplementary Table 5 shows an example of two couples of trials, one belonging to the same pathway and the other not.

The balanced datasets had trials from 124 unique pathways for a total of 2720 couples, while the 32-way 1-shot evaluation validation and test sets consisted of trials from 35 unique pathways each, resulting in 1056 couples for both datasets. Pathways consisting of only 1 trial were used to build couples not belonging to the same pathways. Negative labelled couples were formed also from other trials from different pathways. A scheme of the datasets' composition and origin can be found in Supplementary Table 6.

Trials data

The trial information used for this experiment came from DrugBank. DrugBank contains information parsed from ClinicalTrials.gov. A step of hand curation is performed on the data before entering them to database.

The DrugBank database contains over 142 k trials, out of which only 3277 trials started before 2000. It also contains the information of completed / ongoing trials, and the purpose of the trial.

Model inputs and preprocessing

The inputs of the model were indication, condition, sponsor, phase, country, start date and end date of the trial.

Character-based inputs: character-based inputs considered were indication, condition, country, sponsor. Indication and condition were in the form of lists. The list of text was joined to form the text input. Data augmentation was performed in this case in the form of shuffling the order of the elements of the list.

The preprocessing of the character-based inputs consisted in the removal of stop words. Each input was tokenized at word-level, padded at 1.2 times the maximum length of the training set. For the 1-shot evaluation baselines the input was also 1-hot encoded.

Numerical inputs: the numerical inputs considered in the network were phase, starting date and end date of the trials. These were calculated or inputted and standard scaled.

Additional Inputs: additional inputs were used for the network. These were features preprocessed and concatenated to the absolute difference vector. The inputs were:

- Difference of phases between the two trials
- Days difference between start date of trial 1 and end date of trial 2
- Days difference between start date of trial 2 and end date of trial 1
- Difference between sponsor numbers between trial 1 and trial 2
- Number of common sponsors between the trials
- Difference between the number of countries involved in trial 1 and trial 2
- Number of common countries

These inputs, after they were calculated, were standard scaled on the training dataset.

Neural Network models

The Neural Network models consisted of different branches, depending on the input type (see Supplementary Figure 1 for a scheme of the architectures). These branches contained a single module that encoded trial 1 and trial 2 independently.

Character-based module: Input went through 3 layers of bidirectional (Bi) Long-Short Term Memory (LSTM) (dimension 128, 64, 32 vector size). At the end of the 3 Bi-LSTM layers there was an attention layer, and a fully connected layer (64 nodes).

Numerical branch: Inputs went through a single fully connected layer (64 nodes) and dropout.

After the Siamese modules there was a concatenation layer, which concatenated all embedded inputs from trial 1 and all embedded inputs from trial 2. These concatenation vectors were passed through a layer that provided the absolute difference between the embedded trial 1 and trial 2 vectors.

Additional inputs module: Inputs went through a fully connected (32 nodes) layer and dropout. The output vector was concatenated to the absolute difference vector of trial 1 and 2.

Pre-prediction module: an additional fully connected (64 nodes) and dropout layer that preceded the sigmoid activated prediction layer.

Three models, schematized in Supplementary Figure 1, were used in this experiment:

- A pure Siamese Neural Network model (SNN), consisting of all character-based inputs modules (indication, condition, sponsor, countries) and numerical inputs (phase, start date, end date). No pre-prediction module was added to this architecture
- A hybrid Siamese-Deep Neural Network (SiD NN) which had character-based inputs (indication and condition) and additional inputs (phase difference, difference between start date and end date of the trials, difference between number of sponsors, number of common sponsors, difference between number of countries, number of common countries)
- An Enhanced hybrid Siamese-Deep Neural Network (EnSidNet) with an architecture similar to SiD NN but containing the pre-prediction module

1-shot evaluation baseline models

As baseline models for 1-shot evaluation we used:

1-Nearest Neighbor: calculated as the Euclidean distance between the inputs of the trials. The distance between all inputs was calculated by performing the absolute difference of trial 1 and trial 2, and then summed together.

Random model: couples' similarity was randomly scored.

Metrics

Metrics calculated in this experiment were Precision-Recall Area Under the Curve (PR-AUC) and Area Under Receiver Operating Curve (ROC-AUC), F1-score, precision, recall. Accuracy was an additional metric calculated during the training, on the balanced validation set.

1-shot evaluation assay

A similarity score was assigned to the 32 couples in the batch. If the couple scored most similar was the only couple of trials belonging to the same pathway the batch assay was positive, otherwise negative. The final score was calculated as the percentage of positive hits.

Analysis of the model's feature contribution

To identify the impact of each feature on the overall EnSidNet prediction, a SHAP analysis has been performed on a subset of 10 positive and 10 negative test data.

Prediction pipeline

One of the greatest challenges in implementing a Siamese neural network setting to identify new drug-development pathway (de-novo or completing existing ones) is the number of trials that need to be matched. With more than 140,000 trials, many of which started in the last 20 years, it would be impractical to compare all trials against each other.

The first step of the proposed pipeline was the selection of relevant trials. Trials may be stratified based on the type of sponsor (research institute or pharmaceutical company), the purpose of the trial (e.g. treatment, diagnostic, basic science), phases (phase 4 trials are beyond the scope of this research, so they would be excluded). This first step can reduce the number of trials to match by a factor of 10.

The trials were then divided in buckets based on their therapeutic area. We follow the Medical Dictionary for Regulatory Activities (MedDRA) terminology. The MedDRA System Organ Class (SOC) term was used to represent the therapeutic area. It is rare for trials from the same pathway to include patients affected by pathologies from different MedDRA SOC terms. Dividing the trials into therapeutic area decreased the algorithm complexity. Trials belonging to multiple therapeutic areas were duplicated.

If previous pathways exist for the therapeutic area the algorithm tried to expand them with new trials.

Trial expansion was performed in a setting like 1-shot evaluation. One unmatched trial was compared with 2 trials chosen randomly from all the pathways. The trial was considered to belong to the pathway if the prediction obtained for both trials was higher than a threshold (e.g. 0.8). Corner cases in which trial A and B were matched below the threshold but trial C matched with trial A above the threshold as well as trial B and trial C, were considered a pathway (consisting of trial A, B, and C); this assumption may increase the false positive rate trials in pathway but ensures that all possible clinical trials matching are grouped; the human labelling step would exclude the clinical trials not matching the pathway.

The following step grouped the remaining trials into pathways. To increase the matching probability trials were sorted (for example based on popularity of lead sponsor or condition), then they were divided into batches (in the experiments the batches had 200 trials). Trials within a batch where completely matched. Positive matching was considered for the couples with predictions above a threshold (e.g. 0.8). Matched couples with one trial in common were then grouped into a possible pathway.

To allow grouping of matched trials across batch 1 trial for all possible pathways was matched in an 'all-versus-all' setting, and inter-batch grouping was performed again.

The matching step was repeated 3 times, to ensure the maximum matching of trials.

Once all possible pathways for all therapeutic areas were obtained, the results could be submitted to the human labelers for pathway confirmation.

The false positive couples would be paramount for a second re/training of the algorithm.

Human evaluation of predicted pathways

A subset of the predicted possible pathways across the therapeutic areas (1-3 predicted pathways for each therapeutic area) was sent to human scorers. Trials in the correct pathway kept the drug-development pathway identification number, while trials belonging to a different or new pathway changed the drug-development pathway identification number accordingly. The statistics of the predicted and confirmed pathways can be found in Supplementary Table 4.

B. Supplementary Tables.

NCT ID	Intervention	Condition	Phase	Sponsor	Lead Sponsor	Countries	Date (dd/mm/yy)	
							Start	End
NCT02632708	cytarabine, AG-221, mitoxantrone, daunorubicin, etoposide, idarubicin, AG-120	Newly Diagnosed Acute Myeloid Leukemia (AML), AML Arising From Myelodysplastic Syndrome (MDS), AML Arising From Antecedent Hematologic Disorder (AHD), AML Arising After Exposure to Genotoxic Injury, Untreated AML	1	Agios Pharmaceuticals, Inc., Celgene Corporation	Agios Pharmaceuticals, Inc.	Germany, Netherlands, United States	31/12/15	1/7/23
NCT02073994	AG-120	Cholangiocarcinomas, Gliomas, Chondrosarcomas, Other Advanced Solid Tumors	1	Agios Pharmaceuticals, Inc.	Agios Pharmaceuticals, Inc.	France, United States	1/3/14	1/6/21
NCT02489513	[14C]-AG-120	Healthy Volunteers	1	Agios Pharmaceuticals, Inc.	Agios Pharmaceuticals, Inc.	United States	1/6/15	1/10/15
NCT02677922	Azacitidine, AG-120, AG-221	Leukemia Acute Myeloid Leukemia (AML)	2	Celgene	Celgene	Australia, Canada, France, Germany, Italy, Republic of Korea, Netherlands, Portugal, Spain, Switzerland, United Kingdom, United States	3/6/16	31/10/21
NCT02831972	Itraconazole, AG120	Healthy Volunteers	1	Agios Pharmaceuticals, Inc.	Agios Pharmaceuticals, Inc.	United States	1/6/16	1/10/16
NCT02989857	AG-120 matched placebo, AG-120	Metastatic Cholangiocarcinoma, Advanced Cholangiocarcinoma	3	Agios Pharmaceuticals, Inc.	Agios Pharmaceuticals, Inc.	United States	1/1/17	1/8/20

Supplementary Table 1: Example of a drug-development pathway. Different trials are conducted by pharmaceutical companies to obtain proof of safety and efficacy of the drug before submitting the results to regulatory agency for drug approval

	Number of parameters to train	Average training time (seconds/epoch)
SNN	2,074,497	185.525
SiD NN	1,069,473	96.35
EnSidNet	1,079,681	98.175

Supplementary Table 2: Complexity of the models used for the experiment

	FDA Snapshot	Hand-curated
# of pathways	116	20
# trial/pathway range	1 - 7	1 - 14
25 percentile # trials	1	2
50 percentile # trial	1	4
75 percentile # trials	2	7

Supplementary Table 3: Statistics on the datasets

		Predicted	Confirmed
pathways count		73.000	165.000
trials in pathways	mean	3.616	1.600
	std	5.619	1.258
	min	2.000	1.000
	25 percentile	2.000	1.000
	50 percentile	2.000	1.000
	75 percentile	3.000	2.000
	max	49.000	11.000

Supplementary Table 4: Difference in the distribution of the trial number in predicted vs human checked (confirmed) pathways

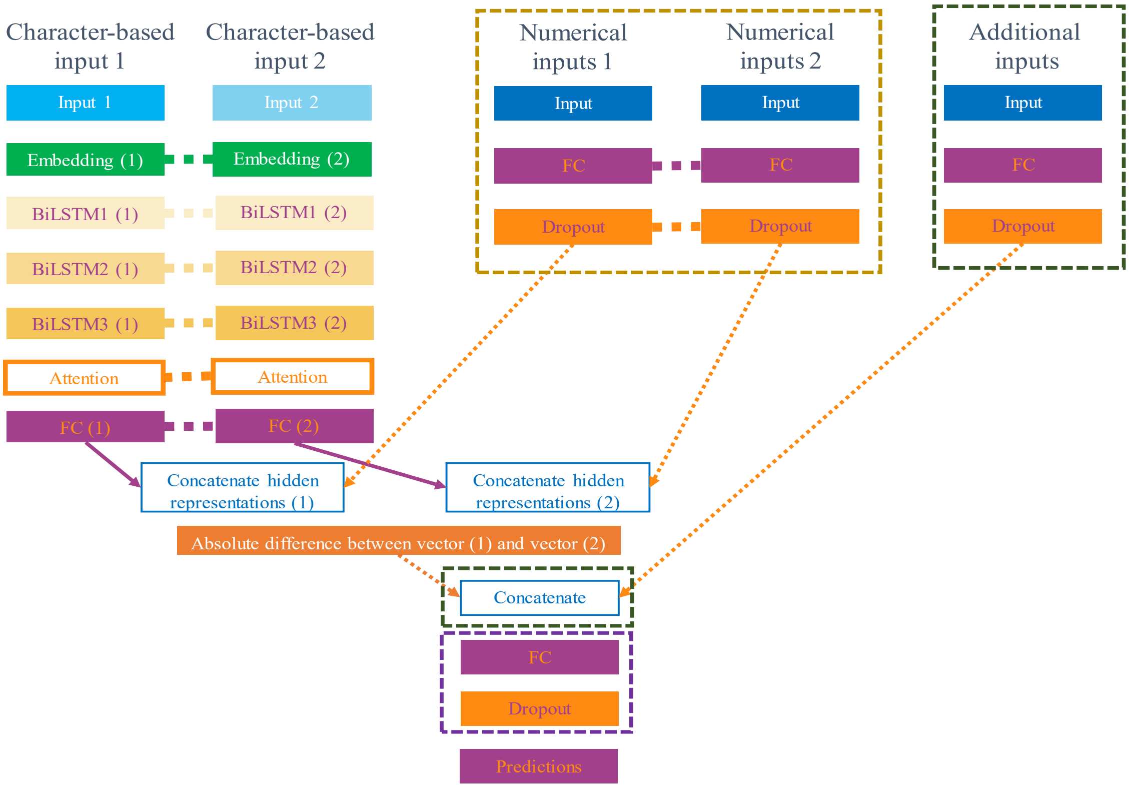
	NCT ID	Intervention	Condition	Phase	Sponsor	Lead Sponsor	Countries	Date (dd/mm/yy)	
								Start	End
Matched	NCT01340872	ST10-021, Placebo Comparator	Ulcerative Colitis, Iron Deficiency Anemia (IDA), Inflammatory Bowel Diseases (IBD)	3	Shield Therapeutics	Shield Therapeutics	Austria, United Kingdom	1/8/11	1/10/14
	NCT02968368	Placebo, Ferric maltol	Iron-Deficiency Anemias, Renal Insufficiency, Chronic	3	Shield Therapeutics	Shield Therapeutics	United States	1/12/16	1/8/18
Not Matched	NCT02946463	Eculizumab, Ravulizumab	Paroxysmal Nocturnal Haemoglobinuria (PNH)	3	Alexion Pharmaceuticals	Alexion Pharmaceuticals	France, Japan, Republic of Korea, United States	20/12/16	1/1/23
	NCT01711359	Baricitinib, Baricitinib Placebo, Folic Acid, MTX Placebo, Methotrexate	Rheumatoid Arthritis	3	Eli Lilly and Company	Eli Lilly and Company	Argentina, Austria, Belgium, Brazil, Canada, Germany, Greece, India, Italy, Japan, Republic of Korea, Mexico, Portugal, Puerto Rico, Russian Federation, South Africa, Sweden, United Kingdom, United States	1/11/12	1/8/15

Supplementary Table 5: Example of a trial couple belonging to the same drug-development pathway (NCT01340872 and NCT02968368) and a trial couple belonging to different drug-development pathway (NCT02946463 and NCT01711359)

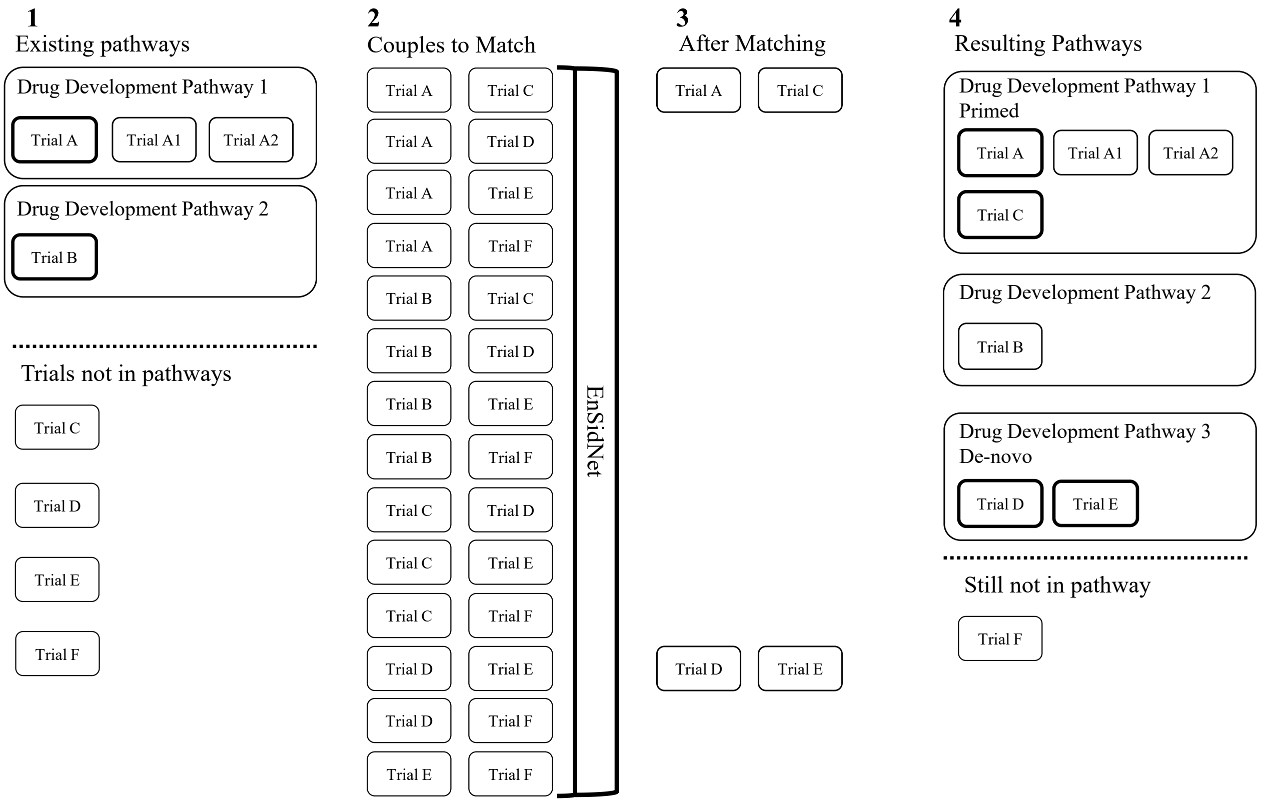
	# total couples	# positive couples	# positive couples’ pathways	# positive couples from snapshot pathways	# positive couples from oncology pathways
Training and validation set	2720	1360	112	101	11
32-way 1-shot validation set	1056	33	33	27	6
32-way 1-shot test set	1056	33	33	29	4

Supplementary Table 6: composition and origin of the datasets

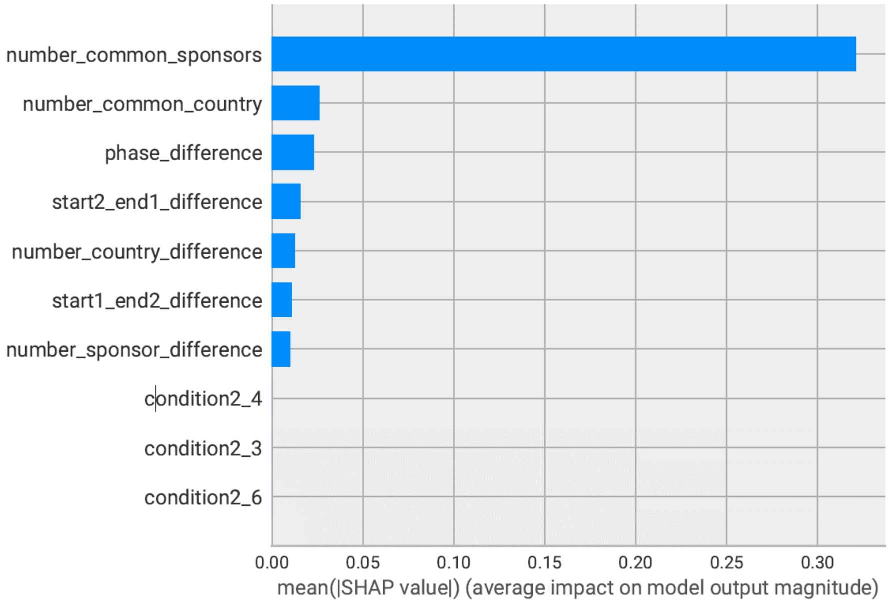
C. Supplementary Figures.



Supplementary Figure 1: Representation of the 3 Neural Network architectures and modules: numerical inputs in gold dashed rectangle (present in the architecture of SNN), additional inputs and a concatenation layer in green dashed rectangle (architecture of SiD NN) and the fully connected layer as last layer before prediction in dark purple dashed rectangle (together with the green dashed module constitute the EnSidNet architecture). BiLSTM = Bidirectional Long-Short Term Memory; FC = Fully connected.



Supplementary Figure 2: Scheme of the matching pipeline. Bold trials in pathways are selected to match to trials not in pathways (here for simplicity only one was selected, in the algorithm proposed they were 2) (**1**). Couples are built (**2**) and matching prediction is given (**3**). Matched trials are combined into existing (primed, e.g. Pathway 1 which included Trial C) or new (de-novo) pathways (e.g. Pathway 3 composed by Trials D and E) (**4**)



Supplementary Figure 3: Feature contribution analysis

636