Privacy-Preserving Knowledge Transfer from Corporate Data to Federative Models

Thierry Hanser
Research Leader

thierry.hanser@lhasalimited.org
The challenge

Model building

Algorithm

Representation

Data
The challenge

Data Model building

Algorithm

Representation

Data

Model building
The challenge

Data

Model building

Algorithm

Representation

Data
The challenge

Model building

Data

Algorithm

Representation
The challenge

Private Knowledge Silos
The challenge

Federating & Sharing knowledge distributed across corporate data without disclosing confidential information
The challenge

Federating & Sharing knowledge distributed across corporate data without disclosing confidential information
Proposed solution: Cronos

Research project to find a solution

Cronos
(Saturn)

Cronos is the patron of harvest. In our case, harvest of knowledge.

GÜNTER, Franz Ignaz 1765-70
Limewood, white painted, height 52 cm
Bayerisches Nationalmuseum, Munich

Cronos is a Lhasa internal project name used only for research purposes.
Proposed solution: Cronos

Teacher - Student approach

Published as a conference paper at ICLR 2017

Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data

Nicolas Papernot*  
Pennsylvania State University  
ngp5056@cse.psu.edu

Martín Abadi  
Google Brain  
abadi@google.com

Úlfar Erlingsson  
Google  
ulfar@google.com

Ian Goodfellow  
Google Brain  
goodfellow@google.com

Kunal Talwar  
Google Brain  
kunal@google.com

Some machine learning algorithms, such as the medical histories of patients, and implicitly store sensitive and private information which cannot be released to third parties. To address this problem, an approach involving strong privacy guarantees has been developed.

Figure 1: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble.
Proposed solution: Cronos

Teacher - Student approach

Sensitive Data (labelled)

Data 1 → Teacher 1
Data 2 → Teacher 2
Data i → Teacher i
Data M → Teacher M

Non Sensitive Data (unlabelled)

Vote

Learn

Non Sensitive Data (labelled)

Space non accessible by adversary

Space accessible by adversary

Learn

Student
Proposed solution: Cronos

Cronos approach

- Member 1
  - Data 1
  - Teacher 1
- Member 2
  - Data 2
  - Teacher 2
- Member i
  - Data i
  - Teacher i
- Member M
  - Data M
  - Teacher M

Non Sensitive Cronos Data (unlabelled)

Consolidation

Learn

Private space

Shared space

Non Sensitive Cronos Data (labelled)

Learn

Student
Proposed solution: Cronos

Cronos research consortium is created with the participation of 8 major pharmaceutical corporates.
The Cronos project

Teachers

Private

Private

Private

Knowledge Transfer

Cronos (350k)

Student

hERG

Student
The Cronos project

---

**Teachers**

- Preissner hERG (4k)
  - Teacher performance
  - Teacher Performance

---

**Knowledge Transfer**

---

**Student**

- Cronos (350k)
  - Student performance
  - Preissner hERG (4k)

---

The Cronos project introduces a new approach to knowledge transfer in education. By leveraging the expertise of experienced teachers (T), the project aims to enhance student performance (S) through personalized learning modules. The project is supported by the Preissner hERG (4k) database, which contains valuable data on hERG blockade, facilitating the design of effective teaching strategies. The integration of private datasets (Private) further enriches the learning environment, allowing for more tailored educational experiences.

---

The Catch-22 of Predicting hERG Blockade Using Publicly Accessible Bioactivity Data

Mikhail B. Stramshetty, Pingfeng Chen, Prashanth Devarakonda, and Robert Preissner

1 Structural Biomembrane Group, Charite - University Medicine Berlin, 10117 Berlin, Germany
2 China Scholarship Council (CSC), Beijing 100034, China
4 DOI: 10.1021/acs.jcim.8b00195
5 Publication Date (Web): May 17, 2018
6 Copyright © 2018 American Chemical Society
7 E-mail: robert.preissner@charite.de

**Abstract**

Drug-induced inhibition of the human ether-á-go-go-related gene (hERG)-encoded potassium ion channels can lead to fatal cardiotoxicity. Several marketed drugs and promising drug candidates were recalled because of this concern. Diverse modeling methods ranging from molecular similarity assessment to quantitative structure-activity relationship analysis employing machine learning techniques have been applied to data sets of varying size and composition (number of blockers and non-blockers). In this study, we highlight the challenges involved in the development of a robust classifier for predicting the hERG end point using bioactivity data extracted from the database.
The Cronos project

Teachers

Student

Knowledge Transfer

Teacher performance
Teacher Performance
Teacher Performance

Preissner hERG (4k)

Perf.(teachers) < Perf. (student) ?

Cronos (350k)

Preissner hERG (4k)

Student performance
The setup

Machine Learning
- SOHN\(^1\) (Self Organising Hypothesis Network)
- Descriptors = Extended Sybyl Atom pairs\(^2\)

Validation
- MCC Matthew Correlation Coefficient (accounts for data set bias)
- Preissner 4502 data points (no ChEMBL overlap)

1. **Self organising hypothesis network: a new approach for representing and structuring SAR knowledge.**
   Thierry Hanser et. al., J. Cheminformatics, 2014, 6, 21
2. **Avoiding hERG-liability in drug design via synergetic combinations of different (Q)SAR methodologies and data sources: a case study in an industrial setting**
The Cronos dataset

Knowledge transporter
- Diversity
- Tractable size
- Homogeneous density
The Cronos dataset

PubChem

100 Millions

Diverse
(too large, not homogeneous)

Random

1 Million

Diverse Tractable
(not homogeneous)
The Cronos dataset

- Retention of diversity without oversampling dense regions
- Optimisation of the chemical space coverage using tiling

Dense

Not homogeneous

Sparse

Cronos
The Cronos dataset

- Retention of diversity without oversampling dense regions
- Optimisation of the chemical space coverage using tiling
The Cronos dataset

1 Million Random Diverse (not homogeneous)

100 Millions Diverse (too large, not homogeneous)

1 Million Diverse Tractable (not homogeneous)

350k Diverse Tractable Homogeneous

PubChem
Label consolidation

Cronos
Unlabeled
(350k)

Teacher 1
S1 L1.1
S2 L2.2
S3 L2.3
...
...
SN L2-N

Teacher 2
L2.1
L2.2
L2.3
...
...
L2-N

Teacher i
Li.1
Li.2
Li.3
...
...
Li-N

Teacher M
LM.1
LM.2
LM.3
...
...
LM-N
Label consolidation

\[ L_S = \frac{\sum_{i=1}^{M} R_i \cdot L_i}{\sum_{i=1}^{M} R_i} \]

Consolidation scheme
The Cronos dataset

PubChem

100 Millions

Diverse (too large, not homogeneous)

Random

1 Million

Diverse Tractable (not homogeneous)

Tiling

350k

Diverse Tractable Homogeneous

Labelling

350k

Diverse Tractable Homogeneous With labels
How much student data is required?

1. Use the best 11k positive student labels
2. Use the best 11k negative student labels

Quality > Quantity when selecting student data
The Cronos dataset

PubChem

100 Millions

Diverse (too large, not homogeneous)

Random

1 Million

Diverse Tractable (not homogeneous)

Tiling

350k

Diverse Tractable Homogeneous

Labelling

Student training set

22K

Diverse Tractable Homogeneous Balanced with strong labels

350k

Diverse Tractable Homogeneous With labels

Distillation
The Cronos dataset

**PubChem**

- 100 Millions
  - Diverse (too large, not homogeneous)

**Random**

- 1 Million
  - Diverse
  - Tractable
  - (not homogeneous)

**Tiling**

- 350k
  - Diverse
  - Tractable
  - Homogeneous

**Student model**

- S

**Learning**

- 22K
  - Diverse
  - Tractable
  - Homogeneous
  - (with good balanced labels)

**Distillation**

- 350k
  - Diverse
  - Tractable
  - Homogeneous
  - With labels

**Labelling**
Evaluation of the student (PoC)

Student compared to teachers (MCC)

Teachers performance range from 0.270 to 0.539
Evaluation of the student (PoC)

Student compared to teachers (MCC)

Average teacher performance = 0.396
Evaluation of the student (PoC)

Student compared to teachers (MCC)

Student outperforms all the teachers on the external test set
### Evaluation of the student (PoC)

- The student could improve precision (52%)
- Very good balanced accuracy (81%)
- Excellent recall (> 91%)
- Excellent negative precision (96%)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>F</th>
<th>D</th>
<th>H</th>
<th>E</th>
<th>C</th>
<th>G</th>
<th>B</th>
<th>&lt;Teacher&gt;</th>
<th>Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCC</td>
<td>0.539</td>
<td>0.325</td>
<td>0.411</td>
<td>0.270</td>
<td>0.364</td>
<td>0.468</td>
<td>0.321</td>
<td>0.472</td>
<td>0.396</td>
<td>0.546</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.529</td>
<td>0.309</td>
<td>0.389</td>
<td>0.256</td>
<td>0.339</td>
<td>0.468</td>
<td>0.255</td>
<td>0.457</td>
<td>0.375</td>
<td>0.499</td>
</tr>
<tr>
<td>BAcc</td>
<td>0.793</td>
<td>0.634</td>
<td>0.731</td>
<td>0.610</td>
<td>0.643</td>
<td>0.737</td>
<td>0.683</td>
<td>0.704</td>
<td>0.692</td>
<td>0.813</td>
</tr>
<tr>
<td>Recall +</td>
<td>0.772</td>
<td>0.353</td>
<td>0.841</td>
<td>0.310</td>
<td>0.350</td>
<td>0.615</td>
<td>0.854</td>
<td>0.473</td>
<td>0.571</td>
<td>0.911</td>
</tr>
<tr>
<td>Recall -</td>
<td>0.814</td>
<td>0.916</td>
<td>0.722</td>
<td>0.910</td>
<td>0.937</td>
<td>0.858</td>
<td>0.511</td>
<td>0.934</td>
<td>0.825</td>
<td>0.714</td>
</tr>
<tr>
<td>Precis +</td>
<td>0.582</td>
<td>0.584</td>
<td>0.472</td>
<td>0.535</td>
<td>0.650</td>
<td>0.593</td>
<td>0.370</td>
<td>0.707</td>
<td>0.562</td>
<td>0.517</td>
</tr>
<tr>
<td>Precis -</td>
<td>0.914</td>
<td>0.808</td>
<td>0.893</td>
<td>0.797</td>
<td>0.811</td>
<td>0.869</td>
<td>0.913</td>
<td>0.841</td>
<td>0.856</td>
<td>0.960</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.587</td>
<td>0.392</td>
<td>0.454</td>
<td>0.328</td>
<td>0.200</td>
<td>0.442</td>
<td>0.307</td>
<td>0.706</td>
<td>0.427</td>
<td>0.774</td>
</tr>
</tbody>
</table>
Merging Cronos data and ChEMBL further increases performance

BAcc = 84% (close to internal validation)
Taking into the confidence of predictions

Selecting the most confident predictions improves the performance. At 80% coverage the balanced accuracy is 88%.
How many teachers are required?

Impact of the number of teachers
(Preissner MCC, student = 6k)

Minimum of 6 members needed
What is the contribution of each teacher?

Member participation:

- A: 94%
- B: 95%
- C: 96%
- D: 97%
- E: 98%
- F: 99%
- G: 100%

Member contribution:

- A: 0.3
- B: 0.15
- C: 0.1
- D: 0.1
- E: 0.25
- F: 0.15
- G: 0.2
- H: 0.1

Strongly collaborative model
What is the impact of individual teachers?

Leave-one-out cross validation of the Student

MCC *

0.516  0.500  0.463  0.455  0.462  0.467  0.478  0.445

All the teachers play an important role

* using a fast student size = 6k
Some students (H,G) are better than their single teacher.
Semi-supervised knowledge transfer provides soft label to the student.
State of the art hERG model

Taking into account confidence leads to very good safety assessment model.
Cronos Student outperforms the Preissner benchmark’s best model (RF/ECFP4). SOHN outperforms the RF. Extended Sybyl atom pairs outperforms ECFP4.

* Preissner corrected for overlap with training test with no ChEMBL overlap
Conclusion

- Effective transfer of knowledge (PoC was successful)
  - Student MCC = 0.546 > Average teacher MCC = 0.396
  - Student outperformed all the teachers in the PoC
- The collaborative model displays state of the art performance for hERG
  - Balanced accuracy = 84-88%
  - Sensitivity = 79-82%
  - Positive precision = 70-80%
- The SOHN algorithm combined with the Extended Sybyl atom pairs provide very promising results
- This was a first round (PoC) and there is a lot of room for optimisation
Perspectives

- Opens an exciting new domain with lots of interesting avenues
- Second round of collaborative research started (Cronos 2)
  - Investigate sparse endpoints
  - Understand and optimise the transfer rate parameters
  - Collaboration is open to new members
- Initiated a consortium to apply the existing transfer methodology (Effiris)
  - Focus on secondary pharmacology targets
  - Consortium is open to new members
- Publication in progress
Acknowledgments

Laura Johnston (coordination with the members and business development)

Ruud Van Deursen (science, cheminformatics and restless computation)

Jeff Plante (science and cheminformatics)

Lorrain Bowe (logistics)

Molecular Informatics Team (cheminformatics platform and support)

Science Team
All the consortium members

✓ Great people
✓ Trust and support
✓ Exciting and motivating scientific discussions
✓ Time and availability
✓ Data
Acknowledgements

...for your attention!