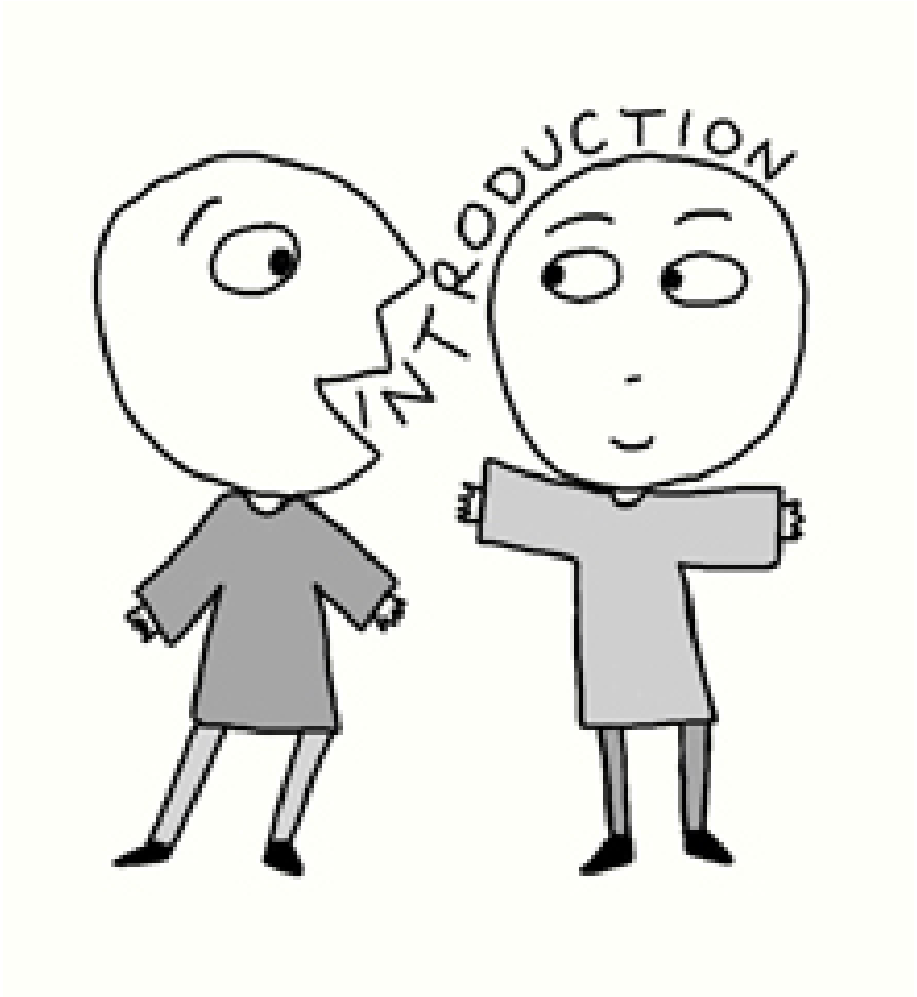
Active Learning of Compounds Activity – Towards Scientifically Sounds Simulation of Drug Candidates Identification

When do we have 90% accuracy?

Wojciech Marian Czarnecki, Stanislaw Jastrzebski, Igor Sieradzki and Sabina Podlewska

Presentation plan

- Introduction: virtual screening and active learning
- Proposed evaluation framework
- Proposed sampling strategy
- Results
- Conclusions

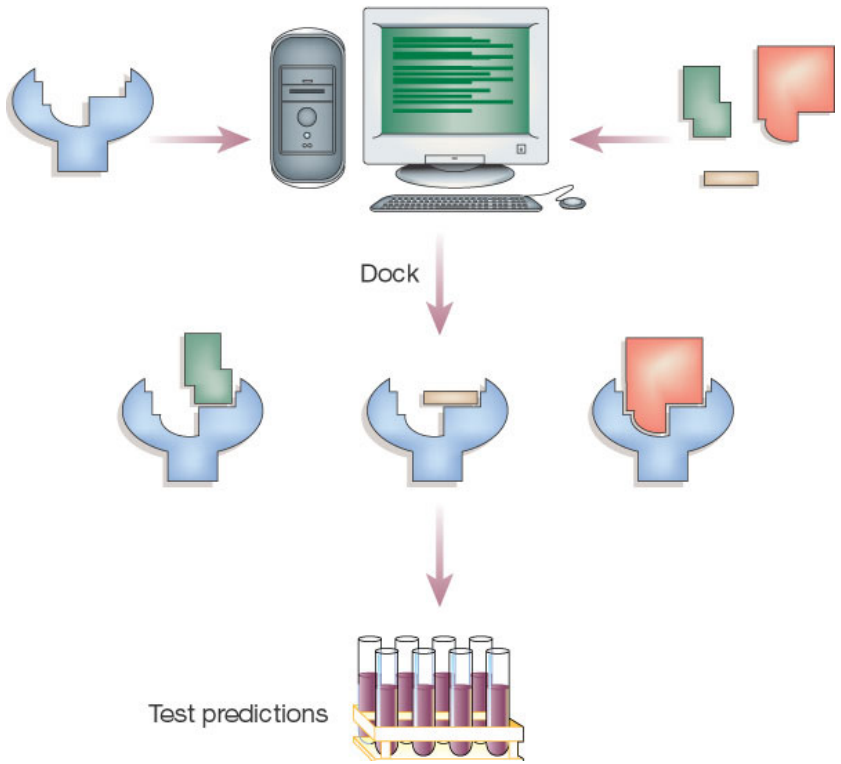

INTRODUCTION

Virtual screening

Virtual screening (VS) is a technique used in drug discovery to filter large body of molecules to identify ones that are likely to bind to a drug target **and will be tested in laboratory later by a chemic**.

Virtual screening

Virtual screening (VS) is a technique used in drug discovery to filter large body of molecules to identify ones that are likely to bind to a drug target **and will be tested in laboratory later by a chemic**.
Most common approaches are structural based (similarity search) and machine learning based.



ML formulation

- Predict if compound will bind to a target in real world, which is framed as a binary classification
- Active compounds are enormously rare but negative results are rarely published (the *positive results bias*). Dataset is not an *uniform* sample of distribution and is **highly skewed**
- Dataset is degenerated, because of the way new drug candidates are created. Most classifiers in naive scenario (binary classification against one target) **degenerate to nearest neighbour**

Active learning

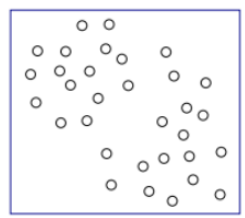
- Unlabeled data is plentiful and cheap: images, speech samples, documents of the web, but **labeling can be hugely expensive**.

Active learning

- Unlabeled data is plentiful and cheap: images, speech samples, documents of the web, but **labeling can be hugely expensive**.
- Sequential process where learner is asking oracle for selected example labels and uses them to retrain

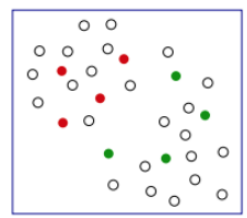
Active learning

- Unlabeled data is plentiful and cheap: images, speech samples, documents of the web, but **labeling can be hugely expensive**.
- Sequential process where learner is asking oracle for selected example labels and uses them to retrain
- Despite active learning success its adoption is still pretty small (only 20% of researchers used it in their projects, as per 2009 survey)



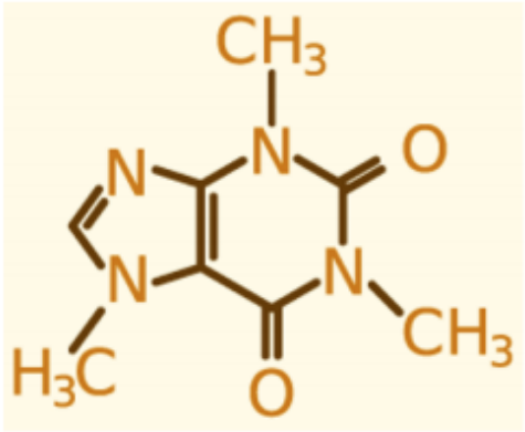
Unlabeled points

Supervised learning



Semisupervised and active learning

Active learning in drug design



- Introduced in 2003 (see Warmuth et al., 2001; Chen et al., 2007)
- Large collection of compounds (catalogs, combinatorial approaches)
- Labeling is extremely expensive (and done in batches)

Previous work

Problems with “classic“ approach to evaluating drug discovery technique in active learning scenario:

- Considers single–instance sampling

Previous work

Problems with “classic“ approach to evaluating drug discovery technique in active learning scenario:

- Considers single–instance sampling
- Assumes “well–behaved“ dataset

Previous work

Problems with “classic“ approach to evaluating drug discovery technique in active learning scenario:

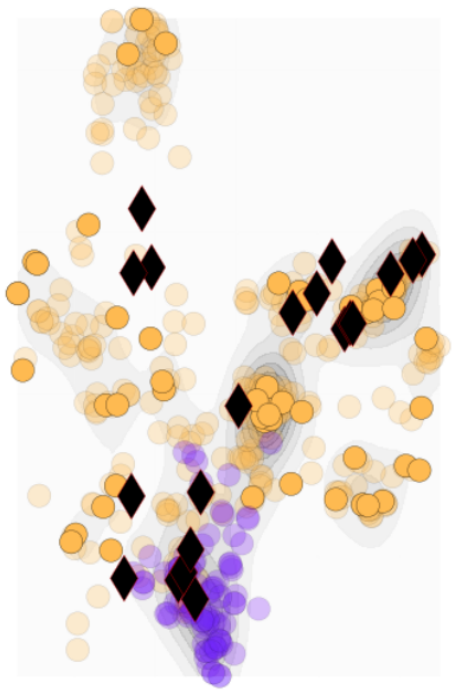
- Considers single–instance sampling
- Assumes “well–behaved“ dataset

$\Rightarrow$ Doesn't answer following question: *Does given active learning strategy lead to the discovery of new, unknown drug candidate?* .

Proposed approach

1. Evaluate sampling strategy through simulation of active learning procedure with k–batch (not single instance) sampling
2. Find a validation cluster. *Do not start AL simulation from the cluster* and *report performance on it*

Proposed approach (cont.)



Proposed approach

Algorithm 1 Drug-discovery evaluation procedure

procedure RunSimulation(X, Y, k)
    Split data into k folds
    Split into two sets, $\mathcal{U}, \aleph$
    for $i = 1$ **to** k **do**
        Train on seed data from $\mathcal{U} \cap X_{train,i}$
        while data in X_{train} **do**
            Select next *batch* of data
            Retrain on *batch*
            Evaluate on $X_{test,i}$ and $\aleph \cap X_{test,i}$
        end while
    end for
    return Averaged metrics over k folds
end procedure

Quasi–greedy strategy

This approach tries to simultaneously optimize for set diversity and sample fitness by finding a set maximizing:

$$u_C(\mathcal{A}) = (1 - C)\frac{1}{|A|}\sum_{a \in \mathcal{A}} u(a) + C\frac{2}{|A|(|A| - 1)}\sum_{a,b \in \mathcal{A}\times\mathcal{A}} d(a, b). \qquad (1)$$

Usually solved in a greedy manner.

Cluster-based Sorensen-Jaccard strategy

Algorithm 2 Cluster-based Sørensen-Jaccard sampling

1: **procedure** $\mathrm{CSJ}_M(\mathcal{U}, k)$
2: $\quad \mathcal{A} \leftarrow \{\}$
3: $\quad U_1, \ldots, U_M \leftarrow$ **find** M **clusters using Sørensen**$(\mathcal{U})$
4: $\quad$ **for** $i = 1$ **to** M **do**
5: $\quad\quad \mathcal{Q} \leftarrow$ **select** k/M **samples by Quasi-greedy using Jaccard**$(\mathcal{U}_i)$
6: $\quad\quad \mathcal{A} \leftarrow \mathcal{A} \cup \mathcal{Q}$
7: $\quad$ **end for**
8: $\quad$ **return** $\mathcal{A}$
9: **end procedure**

Cluster-based Sorensen-Jaccard strategy

Algorithm 3 Cluster-based Sørensen-Jaccard sampling

procedure $\mathrm{CSJ}_M(\mathcal{U}, k)$
  $\mathcal{A} \leftarrow \{\}$
  $U_1, \ldots, U_M \leftarrow$ **find** M **clusters using Sørensen**$(\mathcal{U})$
  for $i = 1$ **to** M **do**
    $\mathcal{Q} \leftarrow$ **select** k/M **samples by Quasi-greedy using Jaccard**$(\mathcal{U}_i)$
    $\mathcal{A} \leftarrow \mathcal{A} \cup \mathcal{Q}$
  end for
  return $\mathcal{A}$
end procedure

Clustering is performed by running k-means after random projection: $\varphi(x) = [S(x, C_1), \cdots, S(x, C_h)]^T$. (see Czarnecki, 2015)

Evaluation

- MACCSFP, PubchemFP and ExtFP fingerprint
- 6 different proteins $\rightarrow$ 6 binary classification problems
- $k = 20, 50, 100$ (greedy should be increasingly suboptimal)
- SVM with Jaccard kernel
- Uses our proposed framework: behavior is investigated on the test set $\mathcal{U}$, $\aleph$ cluster and on unlabeled part of $\aleph$ cluster

Tested strategies

- passive learner
- greedy uncertainty sampling (baseline)
- randomized runs greedy
- CSJ
- Chen and Krause generalized to the nonlinear scenario

Results

Firstly two metrics calculated on $\mathcal{U}$ are analyzed: **final WAC** (WAC$=\frac{1}{2}\frac{\mathrm{TP}}{\mathrm{TP+FN}}+\frac{1}{2}\frac{\mathrm{TN}}{\mathrm{TN+FP}}$) and **area under the WAC curve**. Score is **average ranking**.

batch size	20	50	100	avg
CSJ_2 sampling	**2.33**	**2.17**	**2.17**	**2.22**
Rand Greedy	**2.33**	3.33	**2.17**	2.61
Chen Krause	2.50	2.33	3.50	2.78
Uncertainty	3.17	3.67	3.17	3.33
Passive	4.67	3.50	4.00	4.06

batch size	20	50	100	avg
CSJ_2 sampling	2.17	**2.17**	**2.00**	2.11
Rand Greedy	**1.33**	**2.17**	**2.00**	**1.83**
Chen Krause	4.00	3.00	3.00	3.33
Uncertainty	3.33	3.33	4.17	3.61
Passive	4.17	4.33	3.83	4.11

Results – cntd.

Same two metrics are calculated on ℵ.

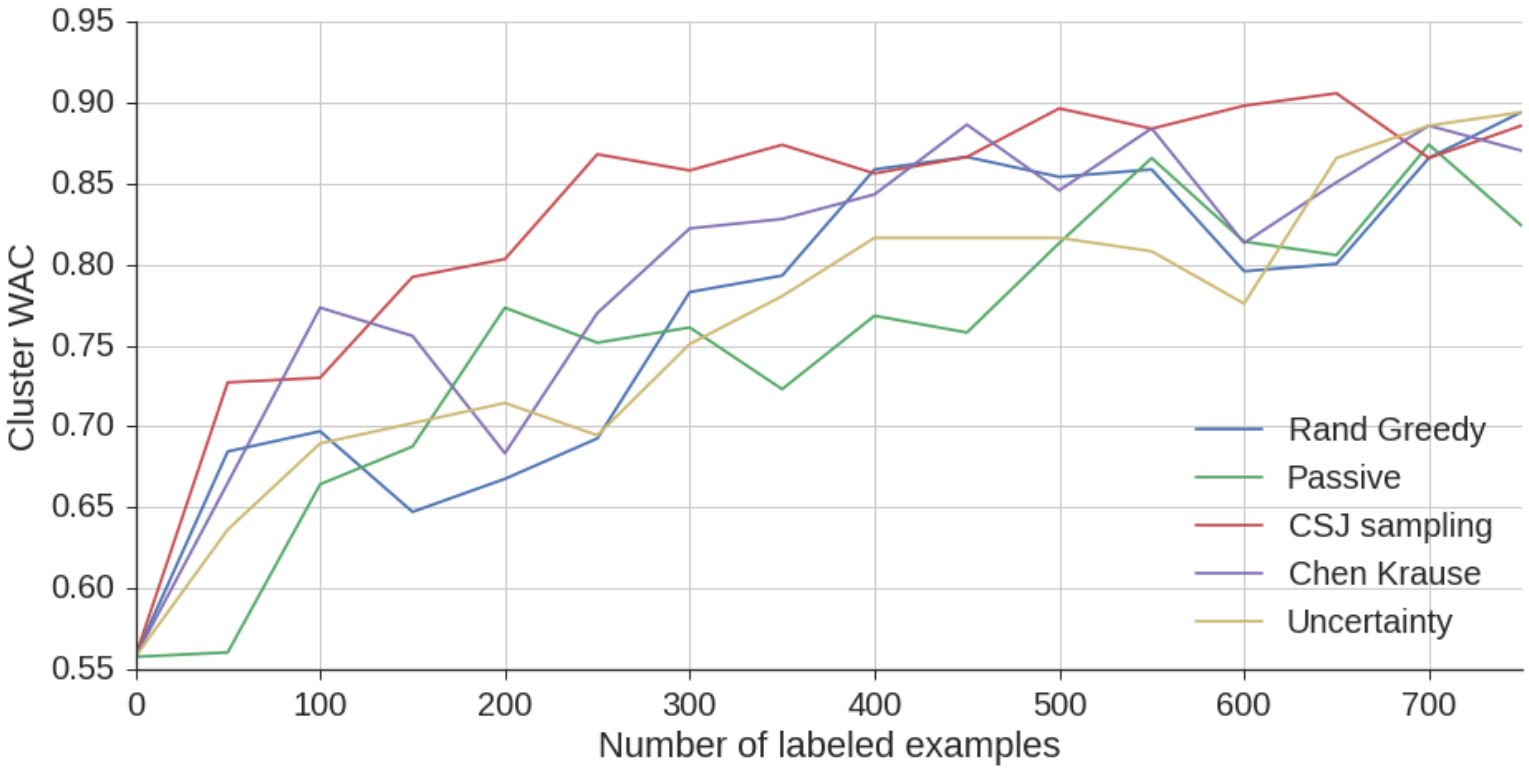
batch size	20	50	100	avg
CSJ_2 sampling	**2.00**	**2.17**	**2.17**	**2.11**
Rand Greedy	2.33	2.50	2.83	2.56
Chen Krause	3.33	3.67	4.50	3.83
Uncertainty	3.83	4.17	2.83	3.61
Passive	3.50	2.50	2.67	2.89

batch size	20	50	100	avg
CSJ_2 sampling	**1.17**	**1.50**	**2.00**	**1.56**
Rand Greedy	2.00	2.17	2.17	2.11
Chen Krause	4.33	4.00	2.83	3.72
Uncertainty	3.33	3.50	3.67	3.50
Passive	4.17	3.83	4.33	4.11

As expected CSJ sampling strategy enforces diversification of sample and thus leading to stronger discovery capabilities.

Results – cntd.

One of the results was that most strategies discover the cluster well, but do not exploit as consistently as CSJ sampling.



Summary

- Most of the drug discovery research is not reporting all the metrics
- We have proposed evaluation framework that should fix it
- We have proposed new sampling strategy that has good results in enforcing diversification (and we have validated that using proposed evaluation strategy)

Future directions

- Test more strategies and fingerprints
- New testing strategy
- Machine learning package alpy in collaboration with Univeristy of Basque (checkout our R package http://r.gmum.net)

Bibliography

Chen, B., Harrison, R., Papadatos, G., Willett, P., Wood, D., Lewell, X., Greenidge, P., and Stiefl, N. (2007). Evaluation of machine-learning methods for ligand-based virtual screening. *Journal of Computer-Aided Molecular Design*, 21(1-3):53–62.

Czarnecki, W. M. (2015). Weighted tanimoto extreme learning machine with case study in drug discovery. *Computational Intelligence Magazine, IEEE*, 10(3):19–29.

Settles, B. (2011). From theories to queries: Active learning in practice. *JMLR Workshop and Conference Proceedings*, 15:1–18.

Warmuth, M. K., Rätsch, G., Mathieson, M., Liao, J., and Lemmen, C. (2001). Active learning in the drug discovery process. In *NIPS*, pages 1449–1456.

Bibliography

Chen, B., Harrison, R., Papadatos, G., Willett, P., Wood, D., Lewell, X., Greenidge, P., and Stiefl, N. (2007). Evaluation of machine-learning methods for ligand-based virtual screening. *Journal of Computer-Aided Molecular Design*, 21(1-3):53–62.

Czarnecki, W. M. (2015). Weighted tanimoto extreme learning machine with case study in drug discovery. *Computational Intelligence Magazine, IEEE*, 10(3):19–29.

Settles, B. (2011). From theories to queries: Active learning in practice. *JMLR Workshop and Conference Proceedings*, 15:1–18.

Warmuth, M. K., Rätsch, G., Mathieson, M., Liao, J., and Lemmen, C. (2001). Active learning in the drug discovery process. In *NIPS*, pages 1449–1456.

Thank you for your attention!

Slides and code @ kudkudak.info