Representation in Machine Learning for Drug Design

Ross D. King,
University of Manchester & Alan Turing Institute
ross.king@manchester.ac.uk

Data Representation

Effective Data Representations

- The key to success in machine learning is the use of effective data representations.
- Almost all machine learning is based on representations that use tuples of attributes, i.e. the data can be put into a single table, with the examples as rows, and the attributes (descriptors) as columns.
- Attributes are normally intrinsic properties of the examples believed to be important: for example if one wished to learn about the effectiveness of a drug, then properties of its molecular structure may be useful attributes.

Notes 1 - Deep Learning

- The most exciting current area of machine learning is that of deep neural networks (DNNs.
- The success of DNNs has been based on their ability to utilize multiple neural network layers, and large amounts of data, to learn how to convert raw input representations (e.g., image pixel values) into richer internal representations that are effective for learning.
- Thanks to this ability to learn effective internal representations, DNNs have succeeded in domains that had previously proved recalcitrant to ML.

Notes 2 – Relational Learning

- The use in machine learning of representations based of tuples of attributes is essence the use of propositional logic.
- Use of the richer, more expressive, language of 1st-order predicate logic is termed Relational Learning.
- Not fashionable now, but in my view, sooner or later relational learning will become essential.
- Used to be considered too inefficient to be practical, but DNNs have moved the bar here!

Talk Plan

- Transformative Machine Learning.
- Relational Machine Learning

Transformative Machine Learning

Ross D. King,
University of Manchester & Alan Turing Institute
ross.king@manchester.ac.uk

Multiple Related Problems are Typical in Science

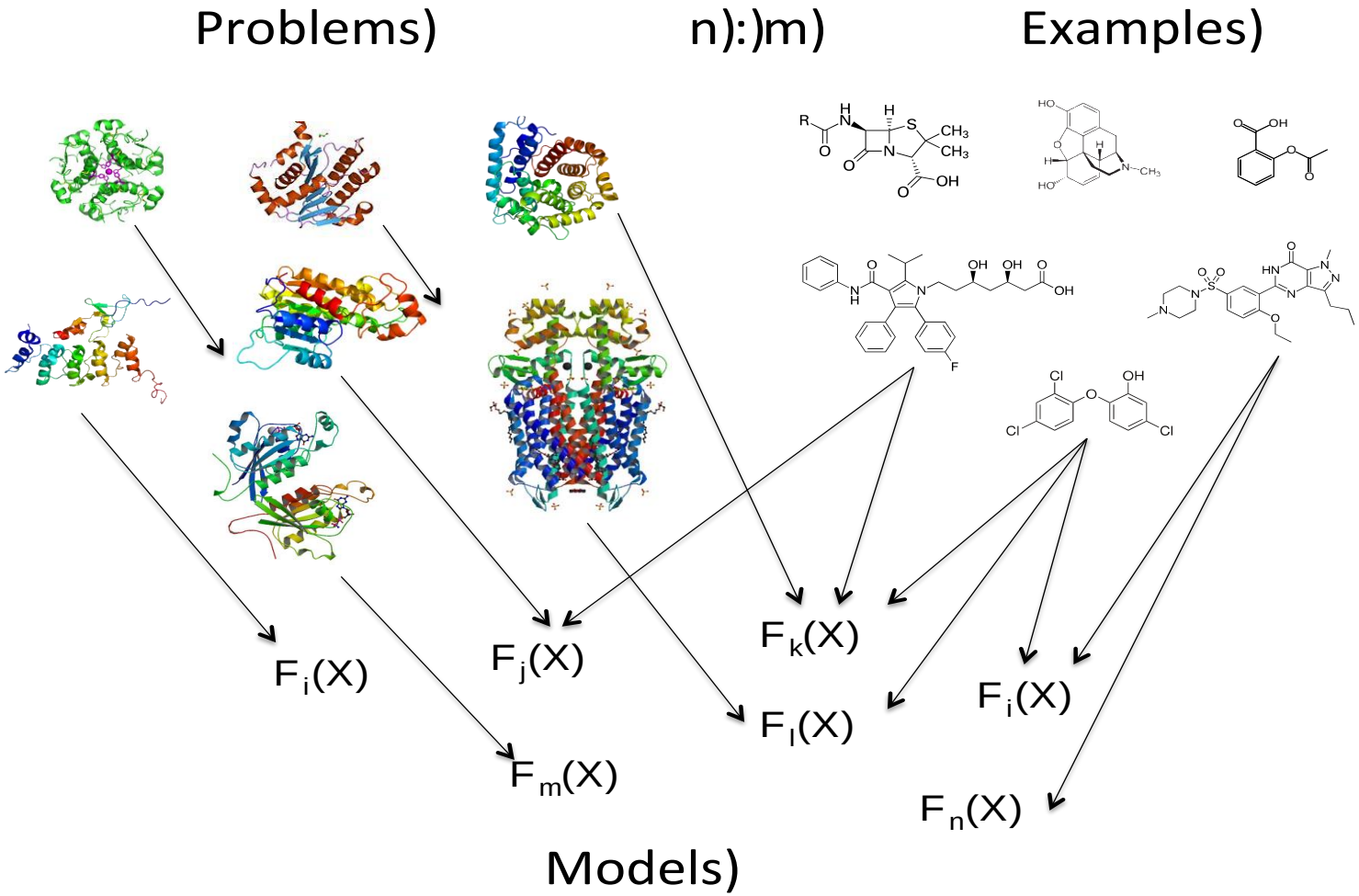
Multi-task learning

- Scientific problems often present themselves as sets of related problems
- Multi-task learning is the branch of machine learning in which related problems (tasks) are learned simultaneously.
- The the aim is to exploiting similarities between the problems to improve performance.
- Problems are learned in parallel using a shared representation, so what is learned from one task can also be used for another problem.

Transfer Learning

- Transfer Learning is closely related to Multi-task learning.
- Information is transferred from a specific source problem to a specific target problem.
- This can be achieved by forcing the target model to be structurally or otherwise similar to the source model(s).
- The success or failure of multi-task or transfer learning often crucially depends on the existence of a good task similarity measure.

Prediction Problems



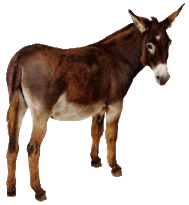
Transformative Learning

The New Idea

- The fundamental new idea is to transform the representation from one based on describing examples using intrinsic properties, to an extrinsic representation based on what other models predict about examples.

- This transformation has the dual advantages of: producing significantly more accurate predictions, and providing explainable models.

Standard Machine Learning



\+ ?

Size	Ears	Cute	Donkey
Big	Big	No	1.0
Small	Big	No	0.3

→ Donkey

\+ ?

Size	Ears	Cute	Kitten
Small	Small	Yes	1.0
Small	Small	No	0.1

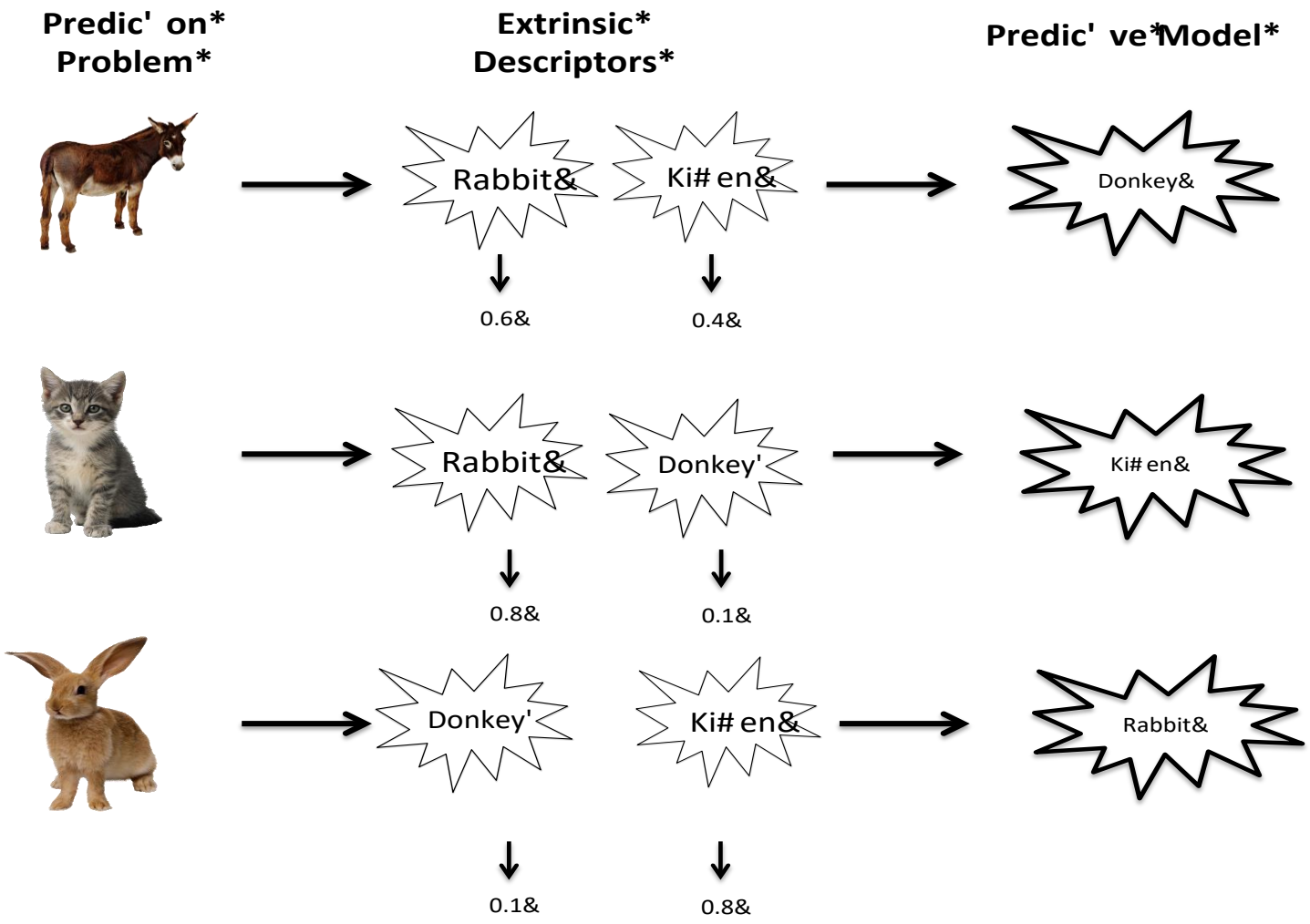
→ Kitten

\+ ?

Size	Ears	Cute	Rabbit
Small	Big	Yes	1.0
Big	Small	Yes	0.2

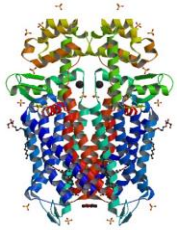
→ Rabbit

Transformative Machine Learning

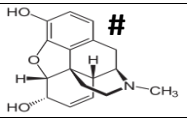
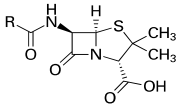


Standard Machine Learning

Predic' on*Problem* **Intrinsic*Descriptors*** **Predic' ve*Model***

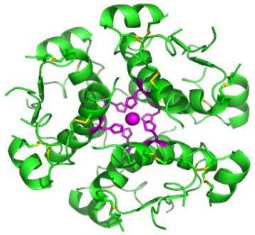


+"

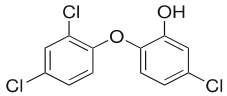
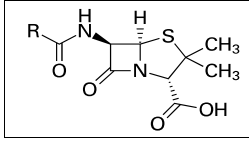
	G1#	G2#	…#	G1024#	Activity#
 #	# 1#	# 0#	# 1#	# 1#	# 0.9#
…#	…#	…#	…#	…#	…#
	# 0#	# 0#	# 1#	# 1#	# 0.3#

→ $F_1(X)$

…" …" …"



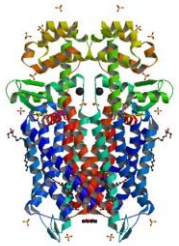
+"

	G1#	G2#	…#	G1024#	Activity#
	# 0#	# 0#	# 1#	# 0#	# 0.5#
…#	…#	…#	…#	…#	…#
	# 1#	# 0#	# 1#	# 1#	# 0.4#

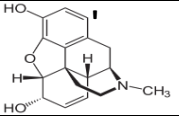
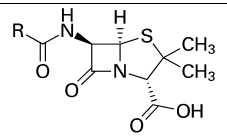
→ $F_n(X)$

Transformative Machine Learning

Predic' on*Problem* **Extrinsic*Descriptors*** **Transforma' ve* Learning* Predic' ve*Model***

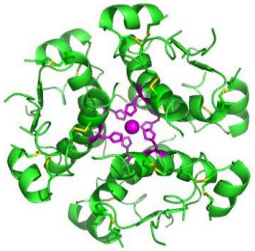


+"

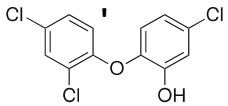
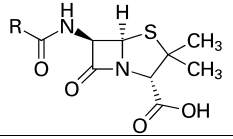
	$\mathcal{F}1$	$\mathcal{F}2$	...	$\mathcal{F}n$	Activity'
	0.3'	0.4'	0.7'	0.9'	0.9'
...'	...'	...'	...'	...'	...'
	0.1'	0.8'	0.7'	0.5'	0.3'

⟶ $T_1(X)$

..." ..." ..."



+"

	$\mathcal{F}1$	$\mathcal{F}2$	...	$\mathcal{F}n$	Activity'
	0.2'	0.5'	0.9'	0.1'	0.5'
...'	...'	...'	...'	...'	...'
	0.5'	0.8'	0.9'	0.1'	0.4'

⟶ $T_n(X)$

Intuition - 1

- The intuition behind Transformative Learning is that instead of describing examples by intrinsic properties, examples are described by what other models predict about them.

- This combining ideas from multi-task learning and transfer learning.

Intuition - 2

- Instead of using a predefined similarity measure to pre-select a set of similar tasks, we project the different tasks into one joint numeric representation.
- Then use a meta-learning algorithm to learn from this new representation how to make accurate predictions for the task at hand.
- Transfer learning has the advantage that it is machine learning method agnostic, and can be applied to improve the performance of many different methods.

The Standard Alternative

- The standard alternative multi-task machine learning approach is to try to learn one large model that encompasses all the problems.
- In many circumstances this can work well.
- However, this standard approach has the disadvantages that if new data, or a new problem is added, then the whole model needs to be relearned.
- Also, neither the relationships between problems, nor the relationships between examples, is made explicit.

Methods

Experimental Setup – Machine Learning Methods

- To in investigate the utility of Transformative Learning we selected four machine learning methods:
  - Random forests (RF).
  - Support-vector machines (SVMs).
  - k-nearest neighbour (KNN).
  - Neural-networks (NN).

- These represent the main families of non-linear Machine Learning methods.

Experimental Setup – Problems

- We applied the four machine learning methods to three typical real-world scientific problems:
  - Drug-design (quantitative structure activity relationship learning - QSAR).
  - Predicting human gene expression (across different tissue types and drug treatments) – LINCS.
  - Meta-machine learning (predicting how well machine learning method will work on problems).

Results

Results

Problem Area	ML Method	Standard NRMSE (RMSE)	Transformed NRMSE (RMSE)	Difference % NRMSE (RMSE)
QSAR				
	RF	0.1643 (0.6511)	0.1478 (0.6316)	10.05 (3.07)
	SVM	0.1693 (0.6720)	0.1522 (0.6490)	10.10 (3.42)
	KNN	0.167 (0.711)	0.171 (0.734)	-2.39 (-3.23)
LINCS				
	RF	(0.0694)	(0.0664)	3.68 (4.32)
	SVM	(0.0692)	(0.0677)	10.9 (2.18)
	KNN			2.92 (-0.28)
	NN	(0.0742)	(0.0707)	-1.15 (4.72)
Meta-ML				
	RF	0.257 (0.1184)	0.124 (0.0526)	51.8 (55.57)
	SVM	0.296 (0.1340)	0.214 (0.0972)	27.8 (27.2)
	KNN	0.290 (0.1330)	0.274 (0.1260)	5.39 (5.40)
	NN	0.323 (0.1480)	0.248 (0.1130)	23.5 (23.7)

XAI -Transformative Learning - 1

- Explainable AI (XAI) is an increasingly important area of machine learning, for in many applications (e.g. medical, financial) there is a necessity to make predictions understandable.

- The understandability of ML models depends on model simplicity, and on how closely the model reflects human concepts.

- The standard theory of human concepts dates back to Aristotle, and is based on the presence of necessary and sufficient conditions.

XAI -Transformative Learning - 2

- The understandability of Transformative learning models is based around the cognitively natural concepts of problem similarity and prototypes.
- This approach to describing human concepts has been been a focus for much recent cognitive science research.
- The Transformative Learning representation is particularly useful if the intrinsic attributes are unsuitable for understandable modelling.

XAI –Human DHFR

	Target ID	Name	Species
1	CHEBL2902	Dihydrofolate reductase	*Lactobacillus casei*
2	CHEBL2014	Nociceptin receptor	*Homo sapiens*
3	CHEMBL2111414	Tyrosine-protein kinase ABL	*Homo sapiens*
4	CHEMBL3048	Nitric-oxide synthase, brain	*Rattus norvegicus*
5	CHEMBL329	Type-1 A angiotensin II receptor	*Rattus norvegicus*
6	CHEMBL4264	24-sterol C-methyltransferase	*Arabidopsis thaliana*
7	CHEMBL5457	Dihydrofolate reductase	*Mycobacterium avium*
8	CHEMBL5372	Methionyl-tRNA synthetase	*Staphylococcus aureus*
9	CHEMBL1075294	Indoleamine 2,3-dioxygenase 1	*Mus musculus*
10	CHEMBL1741172	60S ribosomal protein L 19-A	*Saccharomyces cerevisiae*

XAI – Drug Design Example- 1

- Standard – Explain activity in terms of molecular features.
- Transfer learning – Explain activity in terms of similarity of related problems.

- As expected other DHFR target models are recognised as being useful in predicting human DHFR activity.

XAI – Drug Design Example- 2

- The other target models selected are of biological interest.
- It is interesting that *H. sapiens* Tyrosine-protein kinase ABL was selected, as it has been empirically shown that it is possible to jointly target *H. sapiens* DHFR and tyrosine kinases.
- It is also intriguing that Tetrahydrobiopterin (BH4) is a required cofactor for nitric oxide synthase, and that DHFR regenerates 7,8-dihydrobiopterin (BH2) into BH4.

Large-scale Modelling

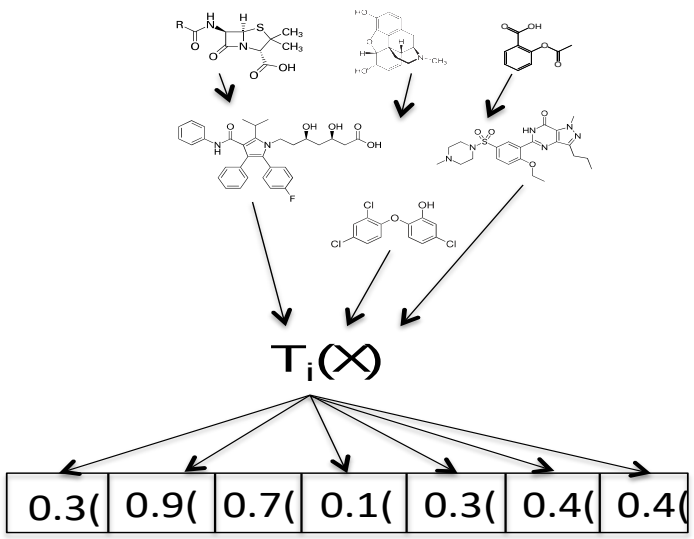
- An important side-product of TL is the large-scale production of predictive models.
- These can be used for clustering, and to make large-scale predictions.

Gene/Protein Clustering

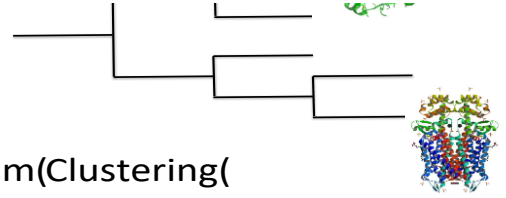
- The standard approach to estimating gene/protein similarity is to estimate evolutionary distance by sequence comparison.
- What is most important in most practical problems is not evolutionary distance, but functional similarity.
- Use of the Transformative Learning predictive models enable genes/proteins to be compared using the accumulated information from millions of empirical experiments.

Clustering

Problem(Profile(

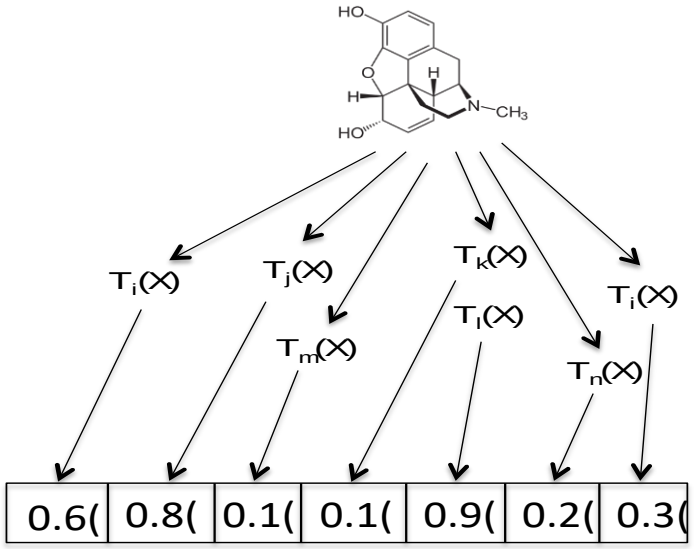


1(Problem((X((n(Examples(

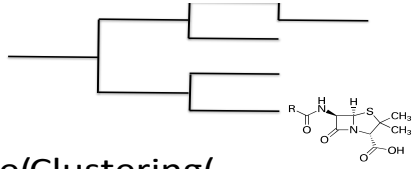


Problem(Clustering(

Example(Profile(



1(Example((X((n(Problems(



Example(Clustering(

Compound Clustering

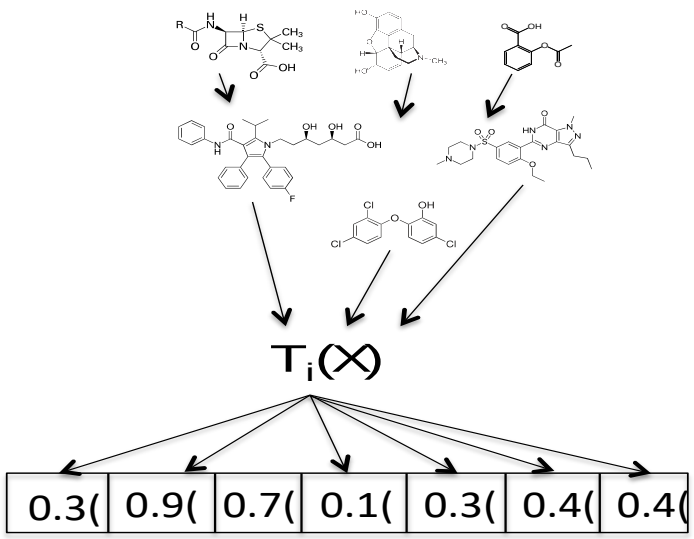
- A fundamental problem in chemoinformatics is the estimation of the similarity of chemical compounds.

- Many different approaches have been applied based on chemical structure.
  - The most commonly used such methods are Tanimoto (Jaccard) coefficient distance between molecular fingerprints, and graph similarity.

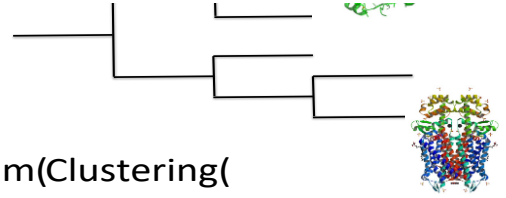
- What is of central interest in most chemoinformatic applications is not structural similarity, but functional similarity.

Clustering

Problem(Profile(

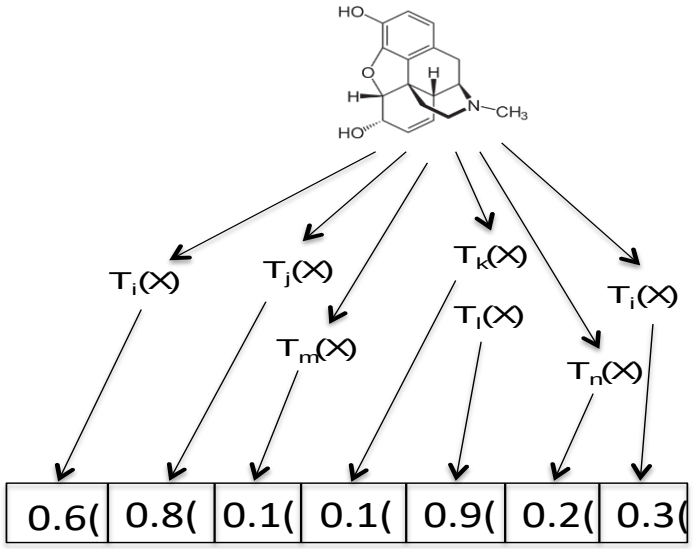


1(Problem((X((n(Examples(

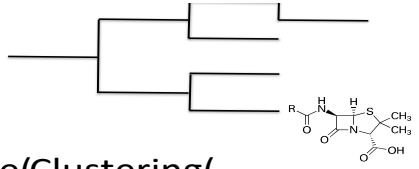


Problem(Clustering(

Example(Profile(



1(Example((X((n(Problems(

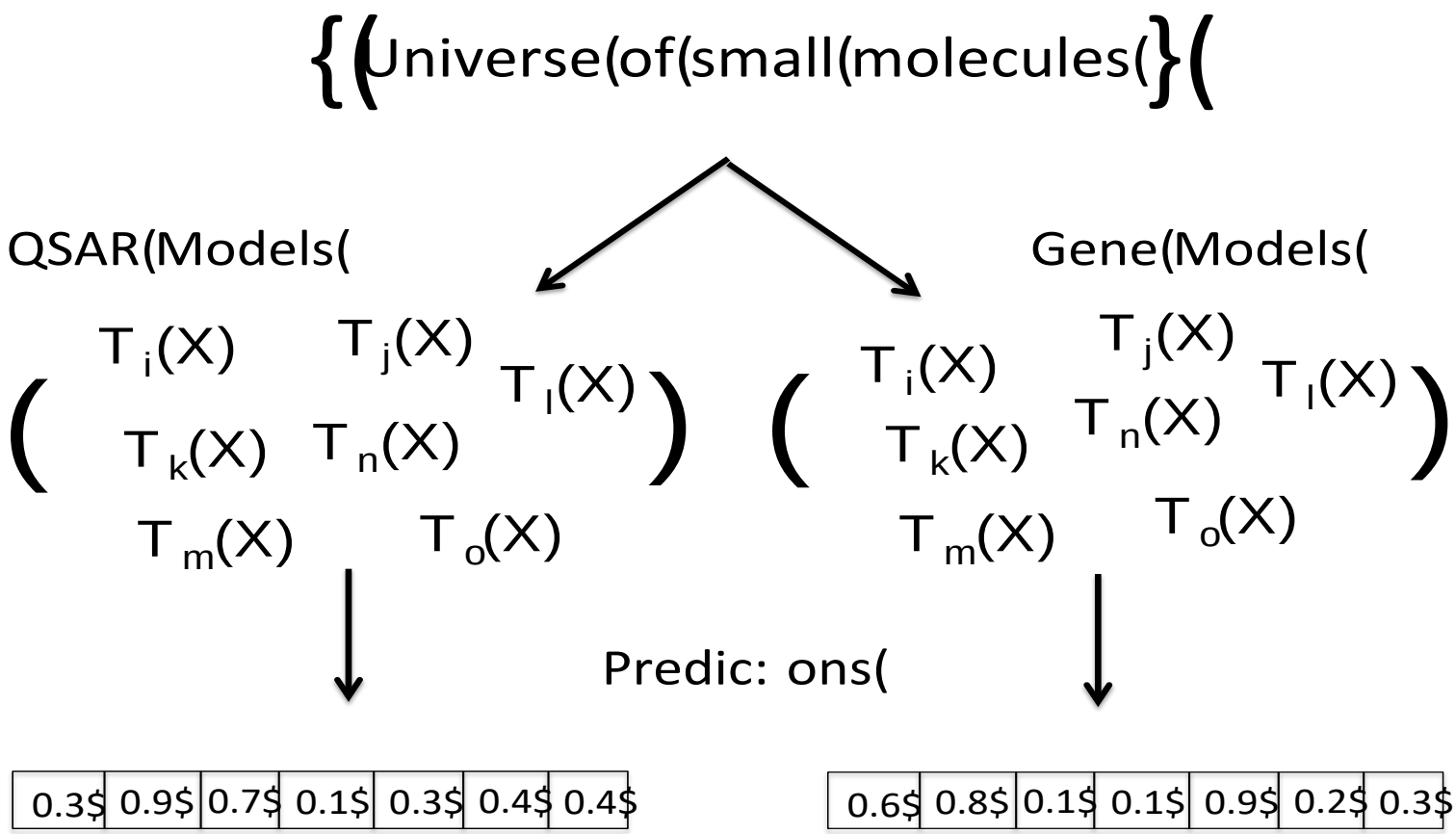


Example(Clustering(

Large-scale Predictions

- The multiple prediction models of Transformative Learning enable large-scale model predictions.
- To illustrate this we have applied our thousands of QSAR target models to predict all the compounds found in ChEMBL
- The universe of possible small-molecules is many many orders of magnitude larger than ChEMBL compounds.
- We therefore plan to bundle up the prediction code and made it freely available.

Large-scale Predictions



FAIR Data

- All the models, predictions, and clustering are fully annotated with meta-data, linked, and freely published on the semantic web.
- To maximize its added-value we follow the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles for publishing digital objects.
- The goal is to facilitate discovery, evaluation, data and knowledge integration and reuse by the community after the publication process.

Discussion

Extensions - 1

- I have presented only the simplest form of Transformative learning. Many other variants are possible with likely improved performance.

- One such approach would be to apply feature selection. The lack of feature selection may explain the K-nn results,

Extensions - 2

- Ensemble Transformative Learning - apply multiple ML methods to obtain better performance than could be achieved by any single ML method alone.
- One possible approach would be to include the standard intrinsic descriptors as well as the Transformative Learning descriptors in the Transformative learning step.
- Stacking could also be used to learn the best ensemble Transformative learning method from different feature selection and ensemble methods.

2nd Order Transformative Learning

- It is natural to extend the idea of Transformative Learning by applying it a second time
  - i.e. to use the predictions from the transformed representation to form a second-order transformed representation.

- As the predictions from the transformed representation are better than the ones from intrinsic representation, learning using second-order transformed representation can be more successful than with the first -order transformed representation.

Conclusions

- We have presented a novel and general Machine Learning representation for sets of related problem.
- The representation has the dual advantages of improving performance, and enabling explainable predictions.
- The fundamental new idea is to transform a standard data representation into an explicit representation based on the predictions of pre-trained models.
- As Machine Learning is increasing being applied to large sets of related problems, we expect Transformative Learning to be of broad general application to scientific problems, and beyond.

Acknowledgments

- Ivan Olier, Oghenejokpeme I. Orhobor, Larisa N. Soldatova, Joaquin Vanschoren.
- This research was partly supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/K030469/1.

Acknowledgments

- Ivan Olier, Oghenejokpeme I. Orhobor, Larisa N. Soldatova, Joaquin Vanschoren.
- This research was partly supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/K030469/1.

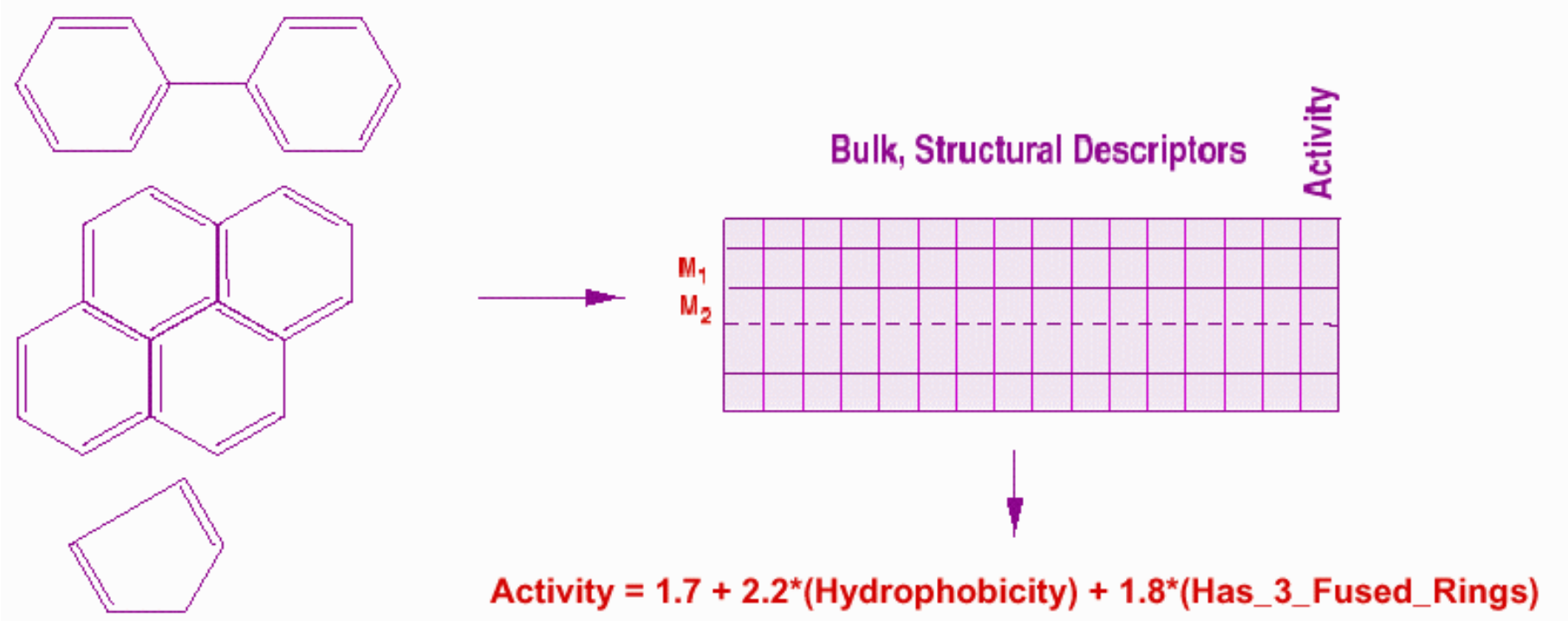
Relational Machine Learning

Ross D. King,
University of Manchester & Alan Turing Institute
ross.king@manchester.ac.uk

Attributes

- Most statistical, neural network, machine learning and data mining methods use attributes to represent examples.
- An attribute is something true about the whole example.
- Example can be put into a single row in a Table.

Classical QSAR



Attributes and chemical structure

- Difficult to represent arbitrary chemical structure using a single table. Either the table grows exponentially, or a compromise must be made in fidelity.

- What should the attributes be?
  - Hansch
  - Boolean fingerprints
  - topological indices
  - voxols

Relational Learning

- Uses richer representation language

  - allows relations and therefore structure

  - based on computer programs (predicate logic)

  - no need to use only a single row of a table

  - easily extended to 3D structure, multiple conformations, etc.

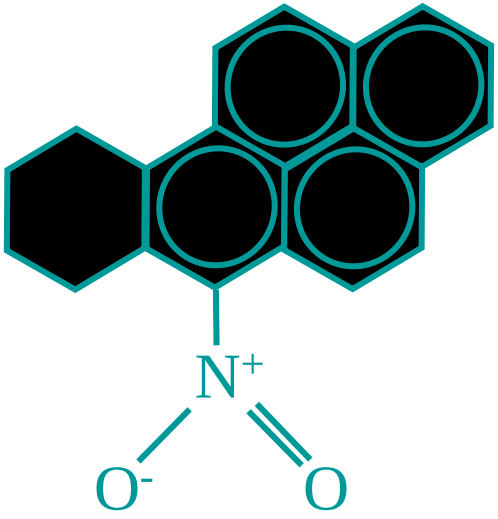
Mutagenesis Example

Example: Predicting Mutagenesis

- Structure-activity relationships of mutagenic aromatic and heteraromatic nitro compounds.
- Aromatic nitro compounds have been found to be both mutagenic and carcinogenic.
- The mutagenesis of these compounds has been studied using the Ames test.

Example Compounds

nitrofurazone



6-nitro-7,8,9,10-tetrahydro[a]pyrene

Hansch Approach 1

- Corwin Hansch was one of the pioneers of QSAR.
- He developed the classical method for QSAR of using linear regression based on describing compounds using attributes such as logP and indicator variables.

Hansch Approach 2

- After much experimentation on the mutagenesis data the following equation was proposed by Hansch and co-workers.

- Log TA98=0.65(±0.16)log P - 2.90(±0.59)log (b.10log P +1)-1.38(±0.25)Lumo+1.88 (±0.39)I1-2.89 (±0.81)Ia-4.15(±0.58)
- n=188, r=0.900, s=0.866, log PO=4.93, log b=-5.48, F1,181=48.6

- This equation gives little structural insight into the QSAR.

Relational Approach

- The compounds are represented using just the atom and bond information.

Relational Approach: Atoms

- The predicate atomm is used, e.g. in compound 127 (3,4,3’-trinitrobiphenyl).
- atom(127, 127_1, c, 22, 0.191).
- States that in compound 127, atom no. 1 is a carbon of type 22 (aromatic carbon in a six membered ring) with a partial charge of 0.191.

Relational Approach: Bonds

- The predicate bond is used, e.g.

  bond(127, 127_1, 127_6, 7).

- States that in compound 127, atom no. 1 and atom no. 6 are connected by a bond of type 7 (aromatic bond).

- There are ~18,300 atom and bond facts in the background knowledge.

High Level Background Knowledge 1

- Relational Learning enables the inclusion of a priori background knowledge.
- The background knowledge consists of computer programs. Arbitrary complex pieces of prior knowledge can be used in the learning process.
- Generally the input programs are in the language PROLOG; but in principle could be any language that can be linked to PROLOG, e.g. FORTRAN.
- It is like using an expert system in the learning process.

High Level Background Knowledge 2

A PROLOG definition of a Methyl group is:

```
methyl(Drug,[Atm0,Atm1,Atm2,Atm3,Atm4]) :
    atom(Drug, Atm0, Type,…,…),          % Link atom
    Type not h,
    atom(Drug, Atm1, c. 10, …),          % Aryl carbon
    atom(Drug, Atm2, h, 3, …),           % Hydrogen
    atom(Drug, Atm3, h, 3, …),           % Hydrogen
    atm(Drug, Atm4, h, 3, …),            % Hydrogen
    bond(Drug, Atm0, Atm1,1),            % Single bond
    bond(Drug, Atm1, Atm2, 1),           % Single bond
    bond(Drug, Atm1, Atm3, 1),           % Single bond
    bond(Drug, Atm1, Atm4, 1),           % Single bond
    Atm2 @> Atm3,
    Atm3@> Atm4.                         % Atm2,3,4 different
```

High Level Background Knowledge 3

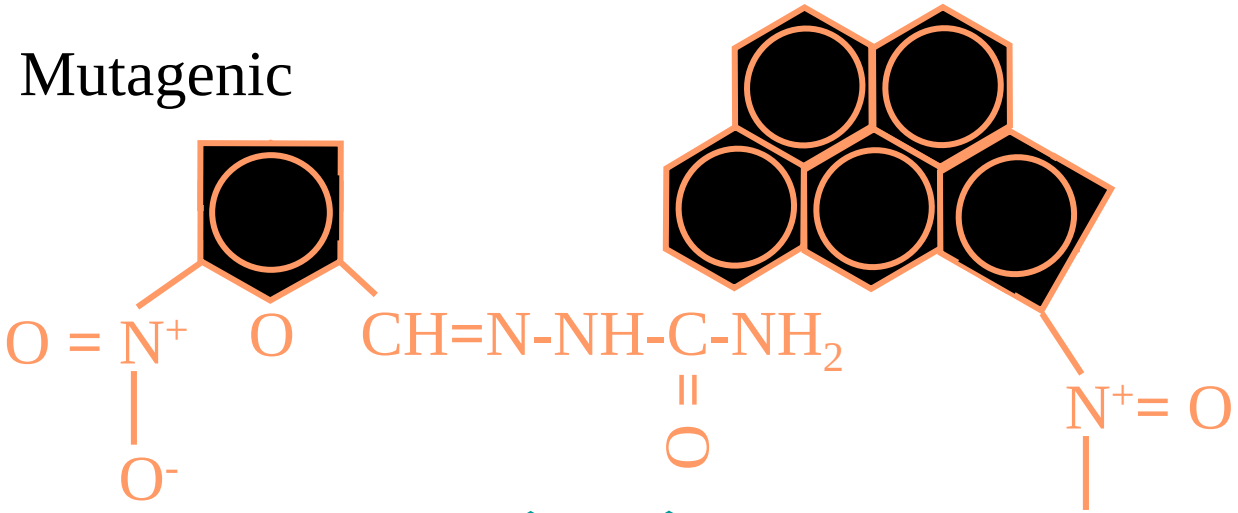
- It is important to appreciate that encoding Progol programs to define concepts is not the same as including them as ‘indicator variables’.
- This method is fully automatic.
- Relational learning can use structural combinations of the chemical groups, e.g. could in theory learn that a structural indicator for activity is diphenylmethane (a benzene single bonded to a carbon atom single bonded to another benzene).

Results 1

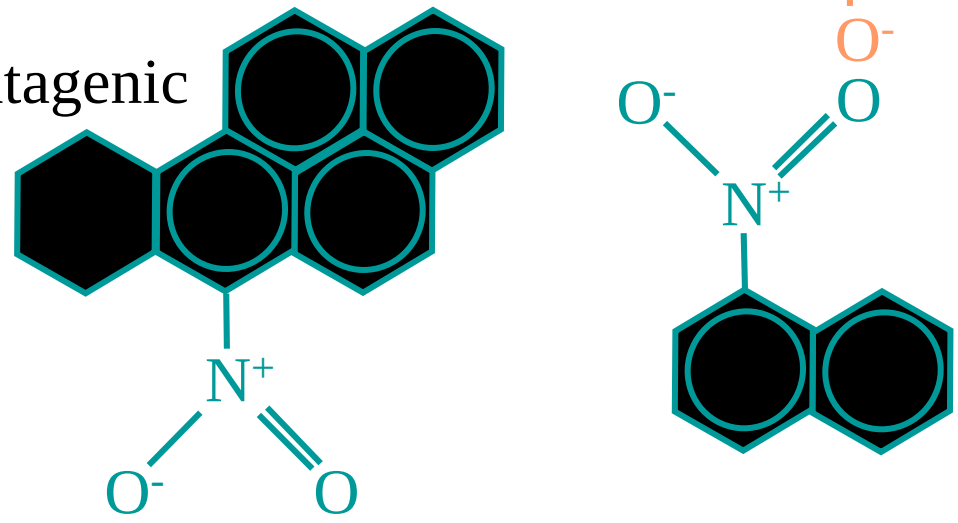
- Hansch split the data into two subsets: 188 compounds which could be explained by a regression equation, and 42 compounds which could not.
- For the 188 compounds a SAR was found that was as good as Hansch.
- For the 42 compounds a SAR was found that was significantly ($P<0.025$) more accurate than linear regression, quadratic regression, or neural nets.

ILP Structural Alert for Mutagenicity

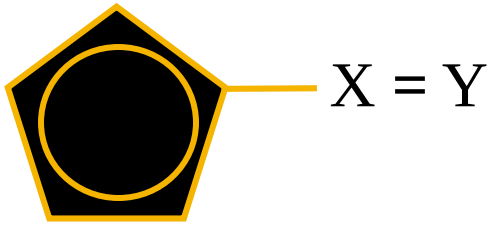
Mutagenic



Non-Mutagenic



Discriminating Pattern



Extensions

- Structure-Based Drug Design
  - Use the 3-dimensional structure of the drug (ligand) and protein.
  - Multiple conformations.

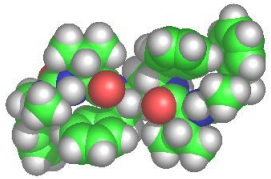
- Quantum Mechanics
  - Molecules are quantum mechanical objects, not classical ones. We wish to represent this quantum mechanical while learning.

- Data Mining
  - Find frequent patterns in sets of compounds.

Three Dimensions

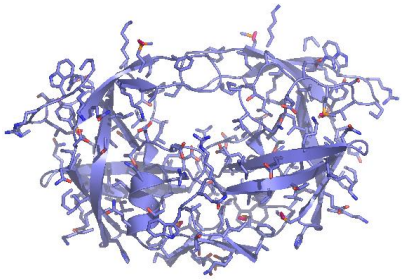
Structure-Based Drug Design

Ligand-based design



Ligands with known activity profile

Structure-based design



Target of known 3D structure

- Features involved in the binding mode to develop new compounds
- Most likely orientation of the ligand in the target (*docking*)
- Estimation of the strength of interaction (*scoring*)

Glycogen phosphorylase *b*

A case study : learning from a series of 3D complexes

Application to 51 structures of the glycogen phosphorylase *b*

- **Water**

Major problem for most of the 3D QSAR techniques

```
:- modeb(*,water(+drug,-conf,-pos)).
```

- **Amino acids**

Computational cost

✓ Properties

✓ Defined active site

```
:- modeb(*,prot_hacc(+drug,-conf,#aa,-pos)).
:- modeb(*,prot_backc2(+drug,-conf,#aa,-pos)).
:- modeb(*,prot_amide(+drug,-conf,#aa,-pos)).
:- modeb(*,prot_lipo_seg(+drug,-conf,#aa,-pos)).
...
```

- **Specific interactions**

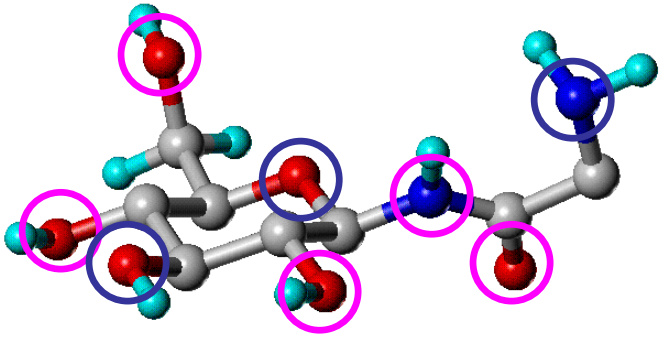
```
:- modeb(*,hb(+drug,-conf,-pos,-pos)).
:- modeb(*,ionic(+drug,-conf,-pos,-pos)).
```

Ligand-based study

51 inhibitors in their crystallized conformation : ideal situation for 3D-QSAR

	R^2_{cv}
CoMFA	**0.46**
ILP	**0.66**
ILP/CoMFA	**0.54**

- ILP > ILP/CoMFA > CoMFA
- Most functional groups involved in the model
- Amide group in that position is related to highly active
- P3 negative contribution



```
log(1/Ki) = 2.41 + 0.84*P1 + 0.95*P2 - 0.40*P3
   P1 : hdon(C),hacc(D),hacc(E),amide(F) ...
   P2 : hdon(C),hacc(D),hacc(E),amide(F) ...
   P3 : hacc(C),hdon(D),hacc(E),ether(F) ...
```

Structure-based study

Comparative study on the information to include to the background knowledge

	R^2_{cv}
Ligand alone	0.66
Ligand+Protein+*hb/4*	0.69
Ligand+*water/4* +*hb/4*	0.75
Ligand+Protein+*water/4* +*hb/4*	0.65
FlexX (scoring function)	0.35

- ✓Need to include receptor information
- ✓ILP outperforms empirical binding energy predictors
- ✓Richness of the background knowledge effect on accuracy
- ✓Closed interpretation of the model as before but
  one localized interaction vs. groups involved in few H-bonds

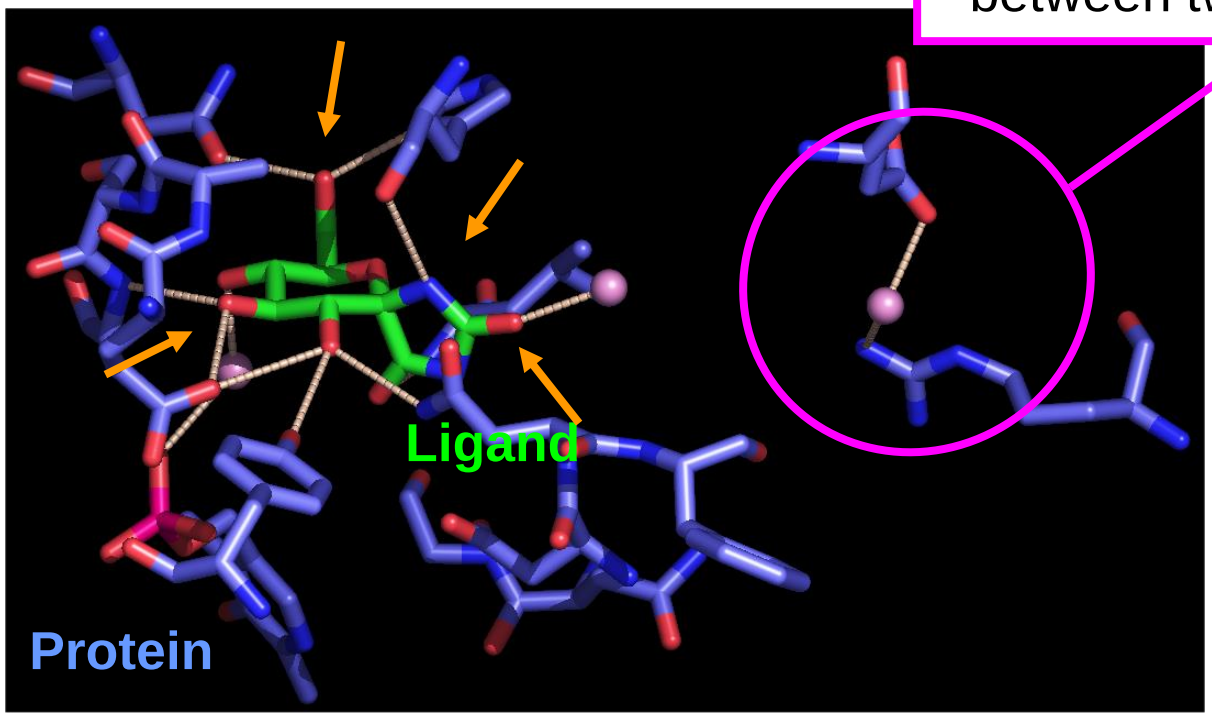
```
log(1/Ki) = 2.43 + 0.76*P1 + 0.91*P2 + 0.35*P3 - 0.49*P4
   P1 : hb(C,D),carbonyl(E),amide(F) ...
   P2 : water(C),alcohol(D),alcohol(E),amide(F) ...
   P3 : water(C),water(D),alcohol(E),amide(F) ...
   P4 : water(C),alcohol(D),methylen(E),equiv_ether(F) ...
```

Mapping onto the active site

3D representation of the pharmacophores proposed by our Relational model

Groups involved in several non-bonded interactions

Stabilisation of the active site via a water bridge between two amino acids



Enot, D. & King, R.D. (2003) PKDD (best paper)

Quantum Mechanics

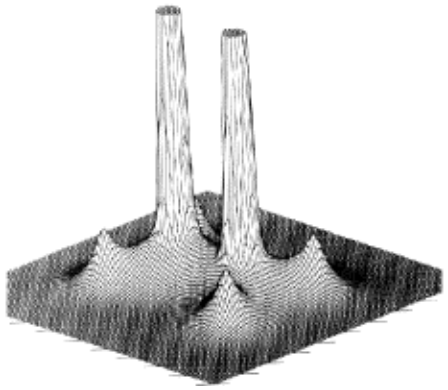
StruQT: Structure representation using quantum topology

Wish-list for molecular structure representations

- Firmly rooted in quantum mechanics - molecules are quantum mechanical objects not classical ones.
- Encode efficiently the electronic structure of the molecule.
- Be related to existing chemical knowledge and concepts.
- Be efficient for machine inference.

Electron density

- Electron density is intuitive for the chemist.
- It is deeply rooted in quantum mechanics.
- CoMFA methods use 3D grid sampling.



A different approach

- The 3D density distribution itself is not well suited for machine representation.
- Use instead the topology of the electron density by inspecting the

Richard F.W. Bader

$$\nabla\rho(r)$$

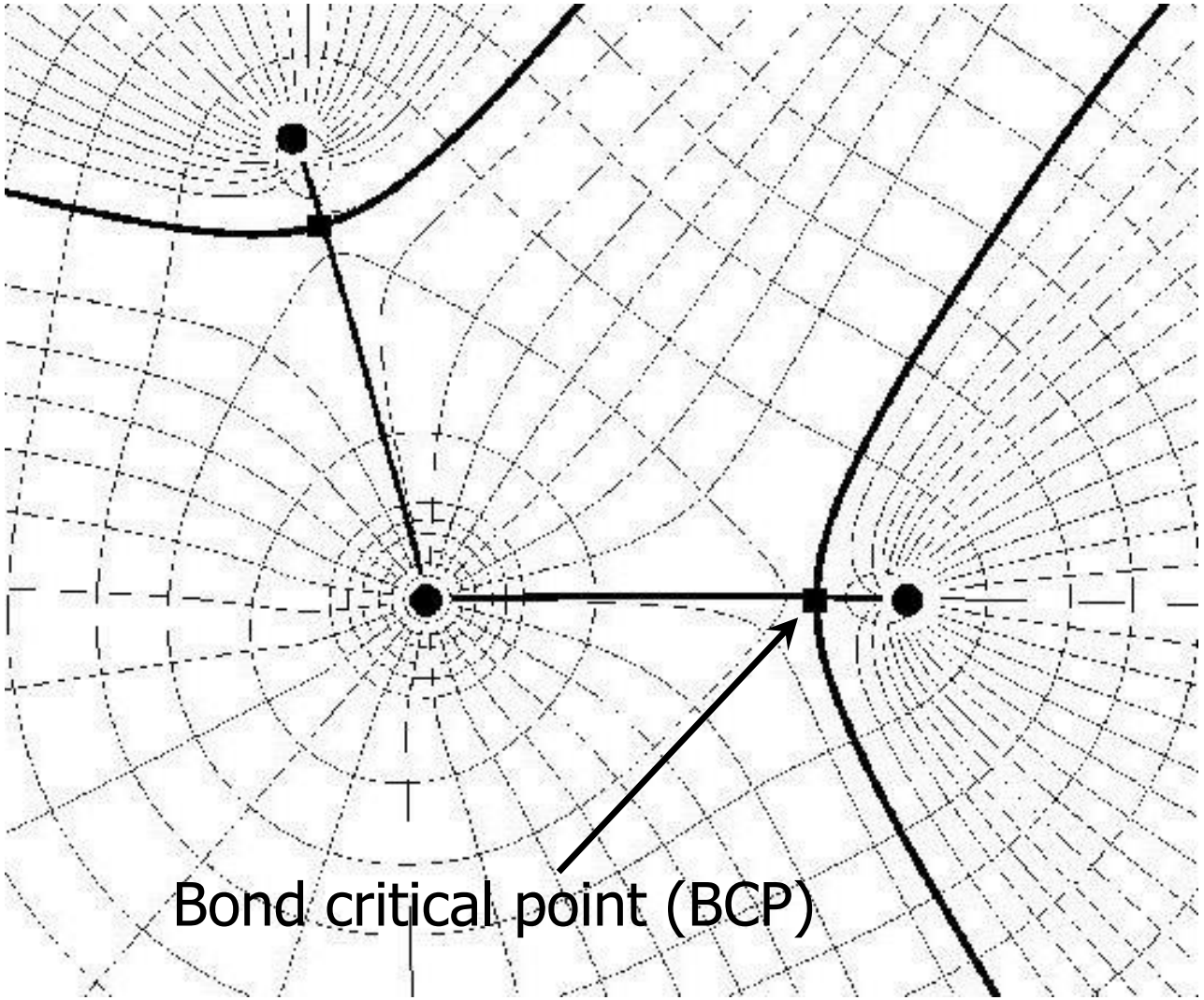
Critical points

- The topology of the electron density is described by a set of critical points (CP) and their properties
- CP:

$$\nabla\rho(r) = 0$$

$$\nabla\rho(r) = u_x \frac{\partial\rho}{\partial x} + u_y \frac{\partial\rho}{\partial y} + u_z \frac{\partial\rho}{\partial z}$$

Gradient paths for water



Example StruQT predicates

- cp(Id,Field,[x,y,z],Hessian) % critical point
- signature(Id,Signature). % signature of Hessian
- rank(Id,Rank). % rank of Hessian
- con_path(Id_1,Id_2) % connectivity of points
- hessian(Id,L_1,L_2,L_3) % eigenvalues of Hessian
- nuc_attr(Id). % nuclear attractor
- bcp(Id) % bond critical point
- rcp(Id). % ring critical point
- ccp(Id). % cage critical point

Data Mining

Data Mining

- The science of finding all frequent “interesting” patterns in data.
- Not to be confused with machine learning or simply searching databases.
- The trick is that if you make a pattern more specific then it can only become less frequent. Therefore can deal with vast amounts of data.

Warmr

- Simplest form: find all frequent combination of items in sets – “basket analysis”. Attribute based.
- Classical example is finding that beer and nappies are often bought together in supermarkets.

- Warmr is a Relational algorithm for finding frequent patterns in logical programs.
- Applied to chemical structures – specific type of program.

Conclusions

- Finding a good representation is central to learning.
- QSAR is in essence a learning task.
- Almost all QSAR methods assume an attribute based representation. This may work well in some circumstances, but it is not a general solution.

- A more appropriate representation is to use *relations*.
- The use of relations implies a Relational based learning approach.
- This has been shown to be effective in many cases.

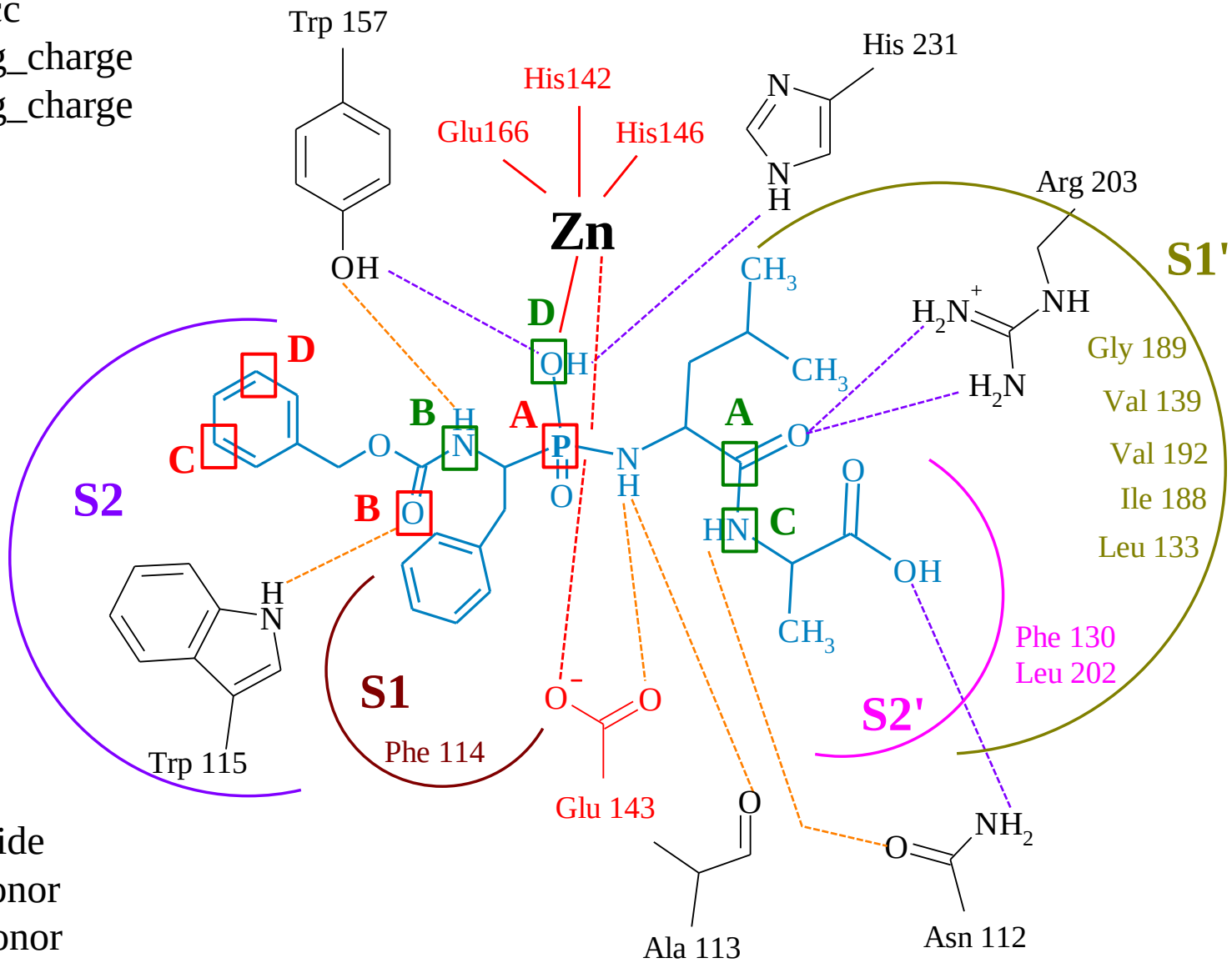
Acknowledgements

- David Enot
- Bjørn Alsberg
- Nathalie Geneste
- Luc Dehaspe

BBSRC, EPSRC

Thermolysin 4TMN

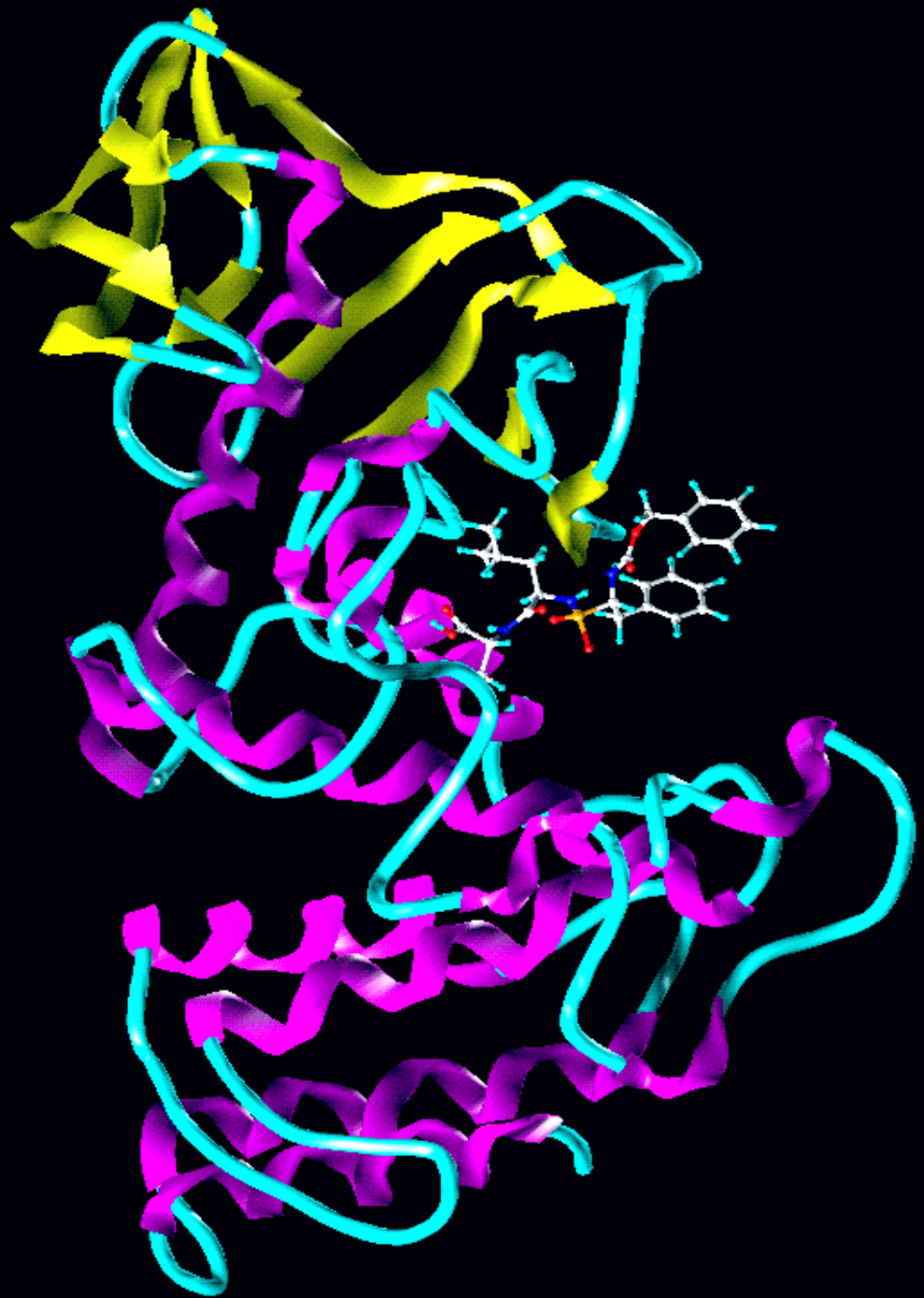
A : phosphorus
B : hacc
C : neg_charge
D : neg_charge

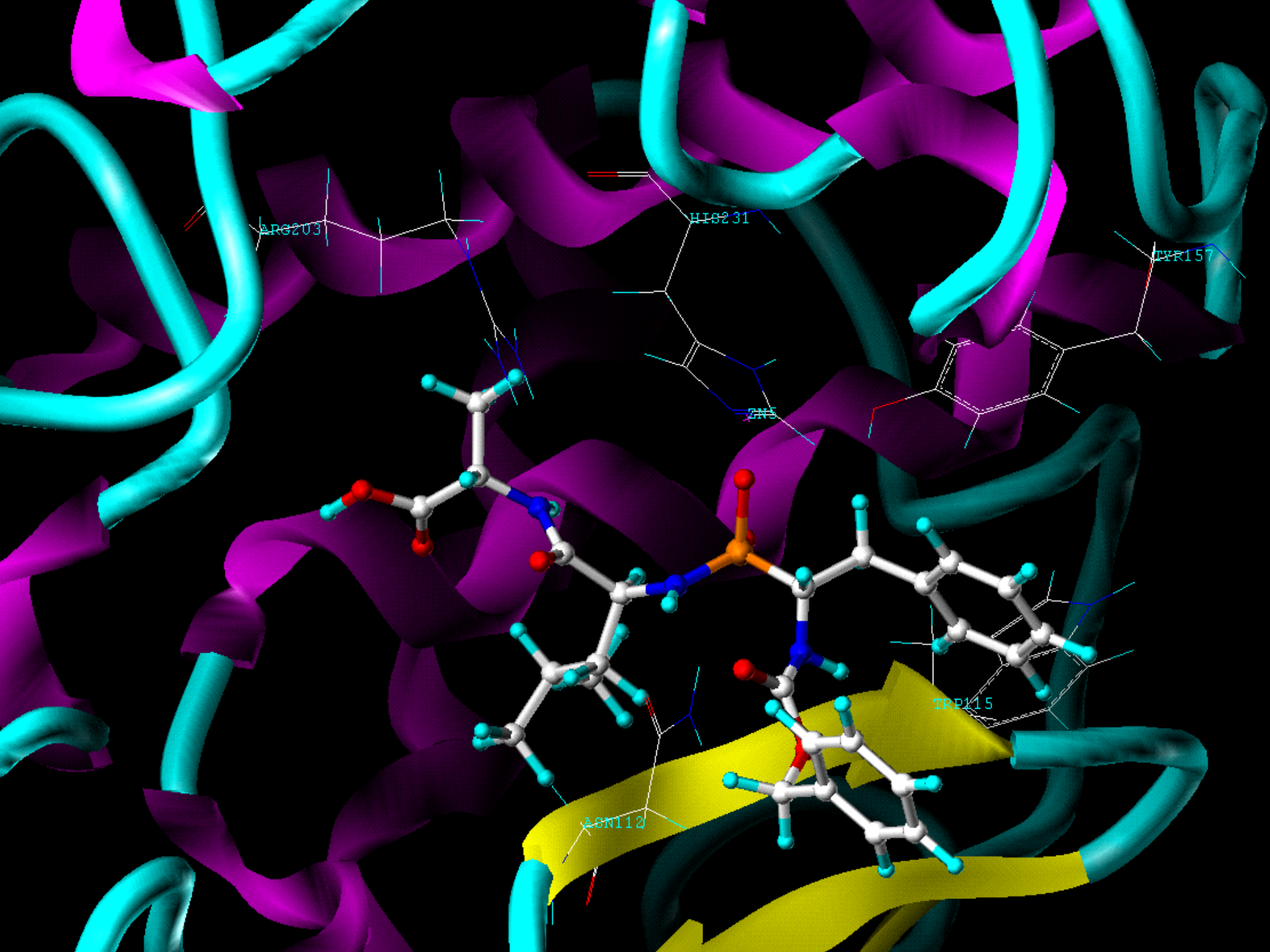


A : amide
B : hdonor
C : hdonor
D : neg_charge

Marchand-Geneste, N., Watson, K.A., Alsberg, B., King, R.D. (2002) *J. Med. Chem.* **45**, 399-409.

4TMN




ARG203
HIS231
TYR157
ZN5
TRP115
ASN112