International Training Program (ITP)
Participation Report

- **Participant Name** : J.B. Brown
- **Academic status** : Ph.D. candidate, 3rd year
- **Home Institute**
  - Laboratory of Biological Network Information (Akutsu Laboratory)
    Kyoto University Institute for Chemical Research; Kyoto, Uji, Gokasho 611-0011
- **Partner Institute**
  - Macromolecular Modelling Group (Knapp Laboratory)
    Freie Universität Berlin; Fabeckstrasse 36a, D-14195, Germany, Berlin-Dahlem
- **Duration of program** : September, 2009 - November, 2009 (approx. 10 weeks)

Contents

1 Collaborative Research Goal **2**
1.1 Background - kernel methods . . . 2
1.2 Separation into two classes . . . 2
1.3 Quantitative Structural Activity-Property Relationships (QSAR/QSPR) . . . 2

2 Research Developments **3**
2.1 Kernel formulations . . . 3
2.1.1 The basic kernel . . . 3
2.1.2 The size-scaled kernel. . . . 5
2.1.3 The rigid kernel. . . . 5
2.1.4 The combination kernel. . . . 7
2.1.5 The stereo-aware combination kernel. . . . 7
2.2 Experimental results from the collaboration . . . 8
2.2.1 How to gauge experimental results . . . 8
2.2.2 Datasets . . . 8
2.2.3 Graphical results . . . 9

3 Non-research international exchange **9**

1 Collaborative Research Goal

1.1 Background - kernel methods

In previous work, J.B. Brown has led the development and testing of a computational method for discerning pairs of stereoisomers using a kernel method [1]. In essence, a kernel method includes a *kernel function* which is a numeric value that represents the similarity between two objects. Kernel functions are designed for applications in a way to extract specific features that the designer feels are important to the task at hand.

With a proper kernel function to represent similarity between objects, pattern analysis algorithms can then be applied for a variety of tasks. One of the most basic pattern analysis problems is the 2-type classification problem. The goal is to predict the class (usually termed "positive" or "negative") of an unseen data point $x \in X$ by using a hypothesis function $h(x)$. X can be a set of chemical structures, photographs, or any other set of objects of a particular class. $h(x)$ is derived using some set of available data of type X.

1.2 Separation into two classes

Recall from linear algebra that the angle θ between two vectors v_1 and v_2 is proportional to their inner product $\langle v_1, v_2 \rangle$:

$$\cos\theta = \frac{\langle v_1, v_2 \rangle}{|v_1||v_2|}$$

In classification problems, the goal is to derive a weight vector w and bias b such that:

$$\langle w, x \rangle + b > 0$$

for points in the "positive" class $((x, y) \mid y = 1)$, and

$$\langle w, x \rangle + b < 0$$

for points in the "negative " class $((x, y) \mid y = -1)$.

Here, the kernel functions mentioned earlier comes into use. Kernel functions have several important properties:

- They satisfy $K(x, y) = \langle \phi(x), \phi(y) \rangle$, which means they are equivalent to transforming $x, y \in X$ to a higher dimensional vectorial space F via a function $\phi : X \to F$.
- They can be applied to non-vectorial data.
- They avoid the problem of inner products on infinite-dimensional vectors.
- They require some algorithm to compute $K(x, y)$.

Perhaps the most important point of the discussion here is the following fact:

Derivation of hypothesis $h(x) = \langle w, x \rangle + b$ using
inner products $\langle x_i, x_j \rangle$ can be rewritten using kernel functions $K(x_i, x_j)$.

1.3 Quantitative Structural Activity-Property Relationships (QSAR/QSPR)

The two-class problem described above can be generalized to predict real values $y \in \Re$. Thanks to this generalization, we can predict target properties of compounds that are expressed as real values. This type of compound analysis is called a Quantitative Structural Activity (Property) Relationship (QSAR/QSPR). This analysis is important in predicting drug efficacy, amongst other properties. An example of differences in topologically identical compounds is given in Figure 1.

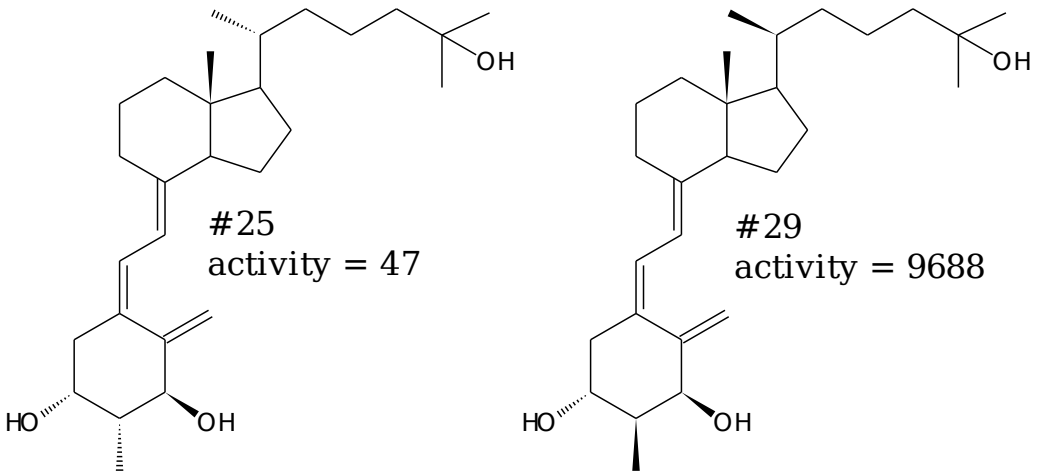


Figure 1: Two vitamin D ligands, identical in topology yet with very different target properties.

2 Research Developments

The general idea of the previous research performed [1] is shown in Figure 2(a). Feature vectors representing counts of molecular fingerprints are built, and the inner product of the feature vectors leads to a measure of similarity. The actual similarity calculation $K(x, y)$ between two compounds x,y is actually done by an alternative method; the details are not relevant to this report.

Though the existing kernel function is generally successful in compound classification and property prediction, it suffers from the requirement of matching labels (atom types) at all nodes in the fingerprint. An example demonstrating the problem is shown in Figures 2(b) and 2(c).

The goal of this international exchange program was to develop a kernel function that uses maximum common substructure but, instead of graph labels, uses electrostatics to quantify the similarity between compounds.

To overcome the problem demonstrated in Figure 2, host advisor Professor Ernst-Walter Knapp suggested use of atomic partial charges as a way to analyze molecules. The most popular method for calculating atomic partial charges is from references [6, 7]. Partial atomic charge maps for the ester and thioester of Figure 2 were created and are shown in Figure 3, where negative partial charge is colored red and positive partial charge is colored blue.

Using the partial charge information, five increasingly sophisticated models were developed.

2.1 Kernel formulations

2.1.1 The basic kernel

The basic kernel simply uses the difference in atomic partial charges between atoms in the maximum common subgraph of two compounds. The concept of why this kernel is useful is given by Figure 3: despite a difference in labelling, the structure of the two motifs are identical, and should be incorporated in similarity quantification.

Definition 1 *M is a maximum common subgraph (or its approximation) between two chemical graphs C_1 and C_2 .*

Definition 2 *PC_a is the partial charge of atom a in compound C .*

Definition 3 *$(a, b) \in M$ is a corresponding match between atom a in C_1 and atom b in C_2 .*

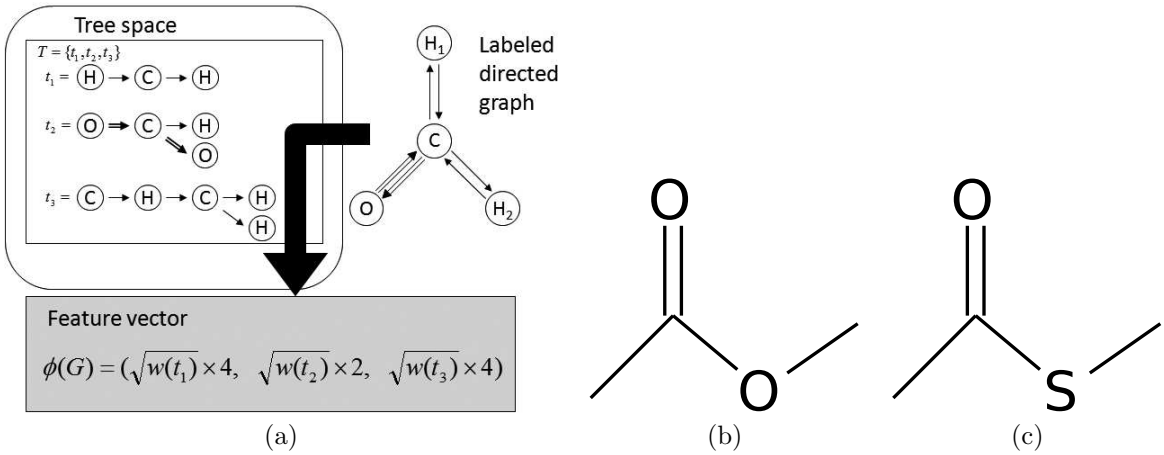


Figure 2: Existing graph kernel methods and their potential problems. (a) In existing methods [1], a molecule is described by counts of a series of tree-like fingerprints. (b-c) The problem with using exact tree patterns. An ester (b) and thioester (c) have the same topology but vary by a single atom label, in which the existing graph methods fail to identify these substructures as similar.

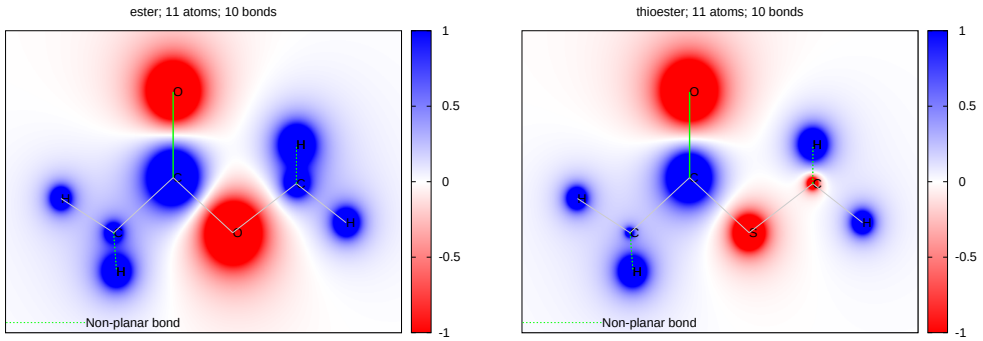


Figure 3: An ester (left) and a thioester (right). Identical topolgies with different atom labels. (Bottom) Atomic partial charge maps, including all hydrogens.

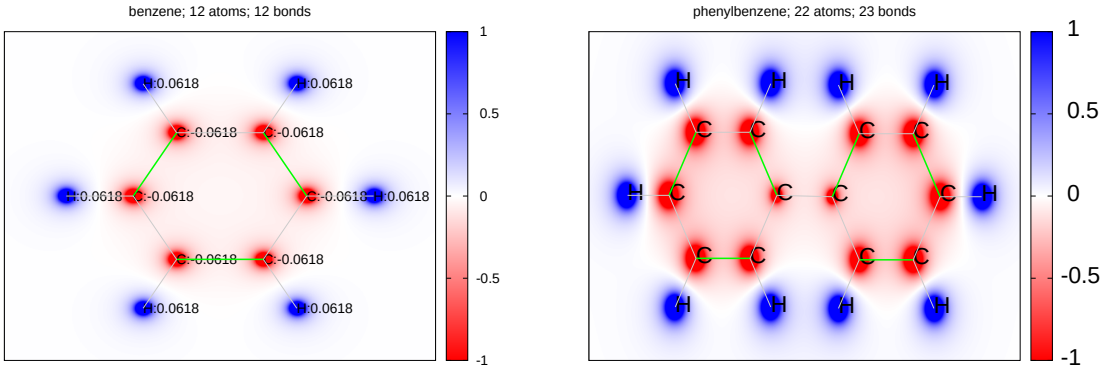


Figure 4: An example case where a molecule scaling coefficient is necessary for similarity calculation: benzene and phenylbenzene.

Formula 1 *Basic kernel definition*

$$K_{\text{basic}}(C_1, C_2) = \sum_{(a,b)\in M} \alpha_1^{-|PC_a - PC_b|} \tag{1}$$

2.1.2 The size-scaled kernel.

It is possible that a small ligand overlaps a much larger ligand or highly specialized and complex drug. In order to reduce the similarity between two compounds when there is a large difference in size, we construct a scaling coefficient that includes the sizes of the original compounds and the size of their maximum common subgraph. A simple scenario where this kernel is useful is given in Figure 4 for benzene and phenylbenzene.

Definition 4 *Let the α_2-scaled overlap between two compounds C_1 and C_2 with maximum common subgraph M, be given by the function*

$$SizeScale(C_1, C_2) = \left(\frac{2|M|}{|C_1| + |C_2|}\right)^{\alpha_2} \quad . \tag{2}$$

Formula 2 *The size-scaled kernel is defined as follows:*

$$K_{\text{size}}(C_1, C_2) = SizeScale(C_1, C_2) * K_{\text{basic}}(C_1, C_2) \tag{3}$$

2.1.3 The rigid kernel.

Figure 5 demonstrates a molecule with a great number of conformers resulting from a long aliphatic. The highly flexible chain may contribute less to the molecule's target property. That is to say, the rigidness of atoms is a key factor in determining their relevance to a similarity calculation, where large target property differences occur when atoms close to the rigid part of a molecule are stereochemically modified. Below we formulate a kernel method to account for rigidity analysis, with an example given in Figure 6.

Definition 5 *The number of rotatable bonds in the path P between an atom a and the nearest rigid atom r:*

$$dist(a, r) = |P = (v_1 = v_a, v_2, ..., v_r)| \tag{4}$$

Definition 6 *An atom scaling effect for rigidity:*

$$RigidScale(a) = \begin{cases} 1 & dist(a,r) \leq 1 \\ \frac{\alpha_3}{dist(a,r)} & dist(a,r) > 1 \end{cases} \tag{5}$$

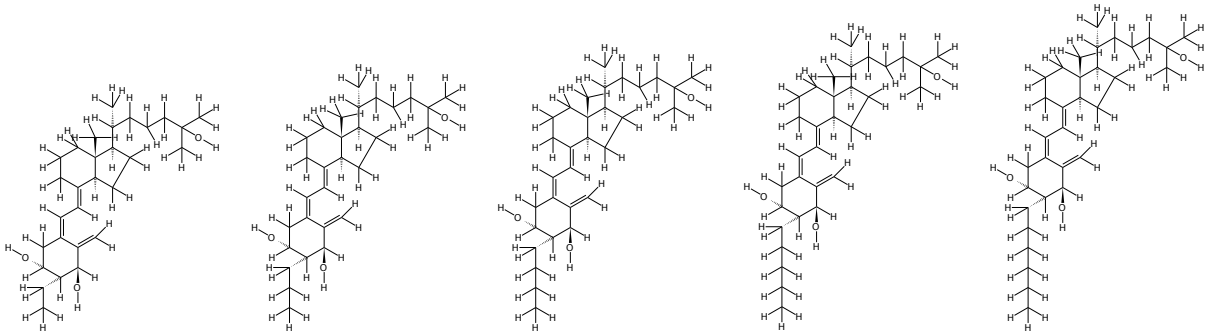


Figure 5: Five vitamin D ligands that vary only by the length of an aliphatic chain extended from the core structure.

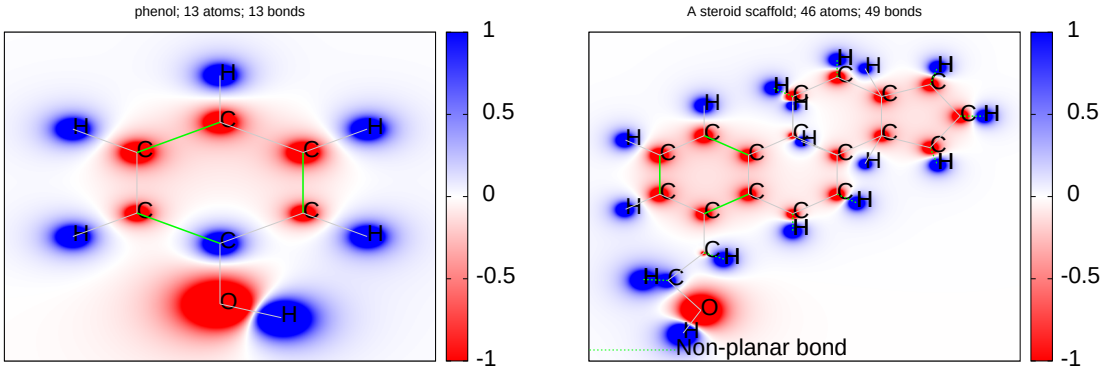


Figure 6: In both molecules, a hydroxyl (-OH) group is present. However, for the larger steroid scaffold molecule, the hydroxyl group is located several rotatable bonds away from the steroid scaffold. Accordingly, the relevance of this hydoxyl fingerprint is reduced by Equation (6).

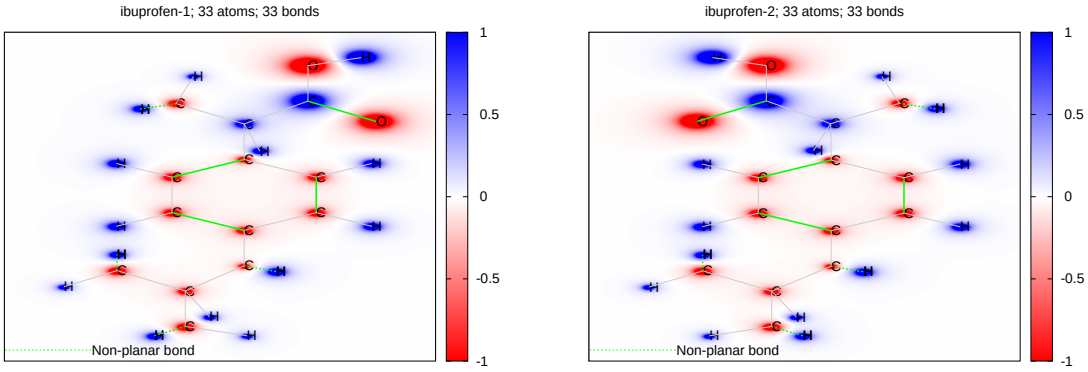


Figure 7: Atomic partial charge density maps of the two crystallized forms of ibuprofen.

Formula 3 *The rigidity kernel*

$$K_{\text{rigid}}(C_1, C_2) = \sum_{(a,b)\in M} RigidScale(a) * RigidScale(b) * \alpha_1^{-|PC_a - PC_b|} \tag{6}$$

2.1.4 The combination kernel.

The combination kernel combines the basic, size, and rigidity kernels:

Formula 4

$$K_{\text{comb}}(C_1, C_2) = SizeScale(C_1, C_2) * K_{\text{rigid}}(C_1, C_2) \tag{7}$$

2.1.5 The stereo-aware combination kernel.

In Figure 7, we present the differences in atomic partial charge density for a pair of chiral compounds. Below, we formulate a kernel function which can handle this type of stereoisomerism, and also generalize the framework to account for *cis-trans* isomerism.

Definition 7 *For a stereocenter or cis-trans carbon atom, define s_1, s_2, s_3, (and s_4 for stereocenters) as the partial charges in the two-dimensional clockwise arrangement of the atoms attached to the carbon.*

Definition 8 *A partial charge difference vector about a chiral stereocenter h is formed by*

$$v_h =< \text{sgn}(s_1 - s_2), \text{sgn}(s_2 - s_3), \text{sgn}(s_3 - s_4) > . \tag{8}$$

Definition 9 *For each end t of a cis-trans double bond chain, the partial charge difference vector will be*

$$v_t =< \text{sgn}(s_1 - s_2), \text{sgn}(s_2 - s_3) > . \tag{9}$$

Next, we need a way to score the directional matching (or mis-matching) of difference vectors.

Definition 10 *For stereoisomer atoms $a_1 \in C_1$ and $a_2 \in C_2$, let v_1 and v_2 be the difference vectors for each stereocenter. Then define a function to weight the importance of matching difference vectors:*

$$RotScore(a_1, a_2) = \begin{cases} \alpha_s & v1 = v2 \\ 0 & v1 \neq v2 \end{cases} , \tag{10}$$

where $\alpha_s \in \{\alpha_h, \alpha_t\}$ is a free parameter set to α_h when processing a pair of chiral stereocenters, and set to α_t when processing a pair of alkene chain terminal carbons. Finally, define a function to use the weighting of Equation (10) when matching stereoisomerism occurs:

Definition 11 *The scaling applied to two stereoisomeric carbons $a_1 \in C_1$ and $a_2 \in C_2$ is defined as*

$$RotScale(a_1, a_2) = \begin{cases} RotScore(a_1, a_2) & stereoType(a_1) = stereoType(a_2) \\ 1 & otherwise \end{cases} , \quad (11)$$

when some function $stereoType(a)$ exists to define whether an atom a is a chiral stereocenter, an end of a cis-trans bond, or neither.

Combining the previous definitions, we now formulate an atomic partial charge kernel function for stereoisomerism:

Formula 5 *The stereochemical-aware combination kernel function:*

$$\begin{aligned} K_{\text{stereoComb}}(C_1, C_2) &= SizeScale(C_1, C_2) * \sum_{(a,b) \in M} \beta(a, b) \\ \beta(a, b) &= RigidScale(a) * RigidScale(b) * RotScale(a, b) * \alpha_1^{-|PC_a - PC_b|} \end{aligned} \quad (12)$$

2.2 Experimental results from the collaboration

2.2.1 How to gauge experimental results

Two measures are used for assessing QSAR prediction performance. The first of these, q^2, is the cross-validated version of the standard residual R^2 and includes the predictive residual sum of squares (PRESS). Let y_i be sample (compound) i's known experimental value (activity level or target property), and let $\hat{y_i}$ be its value output by a predictor during cross-validation. If the known experimental average value is $\bar{y}$, then we calculate q^2 as follows:

$$q^2 = 1 - \frac{\text{PRESS}}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \hat{y_i})^2}{\sum (y_i - \bar{y})^2} \quad . \quad (13)$$

Note that PRESS has no upper bound, and q^2 can take on negative values[4].

The second metric is the correlation R between the predicted and known experimental values for a test dataset after a model has been constructed using the full training dataset[9]. Labeling the average of the predicted values as $\bar{\hat{y}}$, R is defined as

$$R = \frac{\sum (y_i - \bar{y})(\hat{y_i} - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y_i} - \bar{\hat{y}})^2}} \quad . \quad (14)$$

The q^2 and R metrics are standard for assessing prediction performance using LOO-CV and a training-test split. Good prediction performance is signalled when the values of both metrics are close to 1. A perfect predictor would have a $[q^2, R]$ vector with length $\sqrt{2}$.

2.2.2 Datasets

To test the QSARs, two steroid datasets are used independently to build Support Vector Regression models.

The ecdysteroid dataset is a collection of 20-hydroxyecdysone agonists that are involved in the control of ecdysis (shedding) and metamorphosis in arthropods. The EC_{50} value, that is, the effective concentration necessary for 50% of an ecdysteroid to bind to the ecdysteroid receptor and trigger the biological response, is expressed as a numerical value [2, 8, 10, 5, 12]. Dinan *et al.*[5] and Hormann *et al.*[11] have provided the EC_{50} values of 108 ecdysis hormones used in activity prediction research.

Cramer's steroids is a benchmark dataset of 31 steroids used in activity level prediction research* [3, 13, 14]. EC_{50} values represent concentrations required for 50% of a steroid to bind to corticosteroid binding globulin (CBG).

*A number of structures from Cramer *et al.*[3] have been corrected by Silverman[13] and Wagener *et al.*[14]

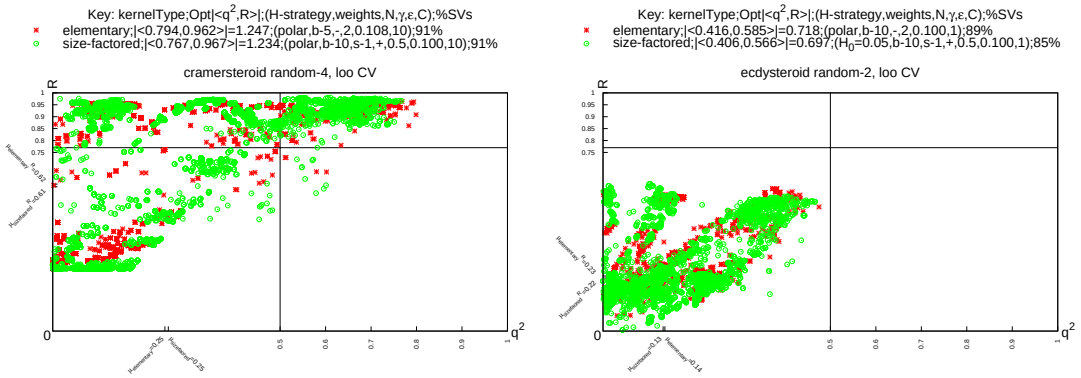


Figure 8: Atomic partial charge experimental results for two types of steroid datasets.

2.2.3 Graphical results

Figure 8 shows the experimental results of QSARs built with the proposed kernel functions combined with Support Vector Regression. Despite their simple formulation, the basic and size-factored kernel functions perform well on the regression tasks. Using the chiral graph kernel formulation, we were unable to reach a model such that $q^2 \geq 0.5$ and $R \geq 0.774$ for the ecdysteroid reference datasets [5, 11], but the two types of proposed kernels that were tested both derived models with sufficient performance (not included in Figure 8, right).

As shown in experimental result figures, less than 100% of the training examples were used as support vectors. This is a positive sign that the atomic partial charge kernels are efficient, because good generalization ability is suggested by the optimals models shown in Figure 8.

This research is still at an early stage, but with promising preliminary results. It is also important to develop a way to use both the graph kernel and the atomic partial charge kernel such that they receive convex weightings depending on graph properties of the dataset being analyzed. That is, efforts should be given to develop a combinined similarity calculation

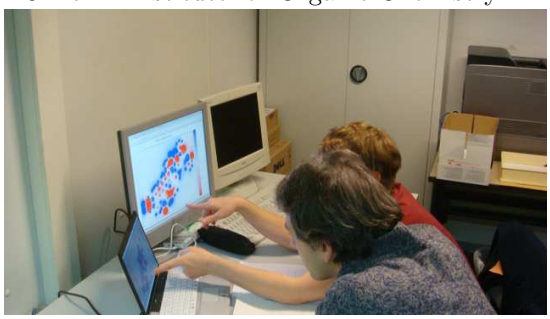
$$\hat{K}(x,y) = \alpha K_{\text{GRAPH}}(x,y) + (1-\alpha) K_{\text{APC}}(x,y) \quad , \quad 0 \leq \alpha \leq 1 \quad . \tag{15}$$

3 Non-research international exchange

Though my time at the Freie Universität Berlin was largely spent at the research laboratory, the members of Knapp Research Laboratory made me feel quite welcome and I enjoyed many other events with the members. Some photos are provided in Figure 9. In addition to new connections in Germany, new connections were established with researchers from Mexico, Italy, Poland, India, Armenia, and Saudi Arabia.

FU-Berlin institute for Organic Chemistry.

Knapp Research Lab individual rooms.



Collaborating on development of kernel methods to analyze steroid structures.

Cooking with students at the research lab to enjoy movies and dinner together.

Enjoying a cruise through Berlin's Spree river with host professor Ernst-Walter Knapp.

Playing a game of football with colleagues in Tiergarten park, located in the center of Berlin.

Figure 9: Photos including research environment and extramural activities.

References

[1] J Brown, T Urata, T Tamura, M Arai, T Kawabata, and T Akutsu. Compound analysis via graph kernels incorporating chirality. Journal of Chemical Information and Modelling, revised:resubmitted, 2009.

[2] C Y Clement, D A Bradbrook, R Lafont, and L Dinan. Assessment of a microplate-based bioassay for the detection of ecdysteroid-like or antiecdysteroid activities. Insect Biochem. Mol. Biol., 23:187–193, 1993.

[3] R D Cramer, D E Patterson, and J D Bunce. Comparitive molecular field analysis (comfa) 1. effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc., 110(18):5959–5967, 1988.

[4] R D Cramer-III. 3D QSAR in Drug Design: Theory, Methods and Applications, chapter The Developing Practice of Comparitive Molecular Field Analysis, pages 443–484. Kluwer Academic Publishers: Dordrecht, The Netherlands, 2008.

[5] L Dinan, R E Hormann, and T Fujimoto. An extensive ecdysteroid comfa. J. Comput.-Aided Mol. Des., 13:185–207, 1999.

[6] J Gasteiger and M Marsili. A new model for calculating atomic charges in molecules. Tetrahedron Letters, 34:3181–3184, 1978.

[7] J Gasteiger and M Marsili. Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges. Tetrahedron, 36:3219–3228, 1980.

[8] A Golbraikh, D Bonchev, and A Tropsha. Novel chirality descriptors derived from molecular topology. J. Chem. Inf. Comput. Sci., 41:147–158, 2001.

[9] A Golbraikh and A Tropsha. Beware of q2! J. Mol. Graphics and Modell., 20:269–276, 2002.

[10] A Golbraikh and A Tropsha. Qsar modeling using chirality descriptors derived from molecular topology. J. Chem. Inf. Comput. Sci., 43:144–154, 2003.

[11] R E Hormann, L Dinan, and P Whiting. Superimposition evaluation of ecdysteroid agonist chemotypes through multidimensional qsar. J. Comput.-Aided Mol. Des., 17:135–153, 2003.

[12] M Ravi, A J Hopfinger, R E Hormann, and L Dinan. 4d-qsar analysis of a set of ecdysteroids a comparison to comfa modeling. J. Chem. Inf. Comput. Sci., 41:1587–1604, 2001.

[13] B D Silverman. The thirty-one benchmark steroids revisited: Comparitive molecular moment analysis (comma) with principal component regression. Quant. Struct.-Act. Relat., 19:237–246, 2000.

[14] M Wagener, J Sadowski, and J Gasteiger. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic ah receptor activity by neural networks. J. Am. Chem. Soc., 117:7769–7775, 1995.