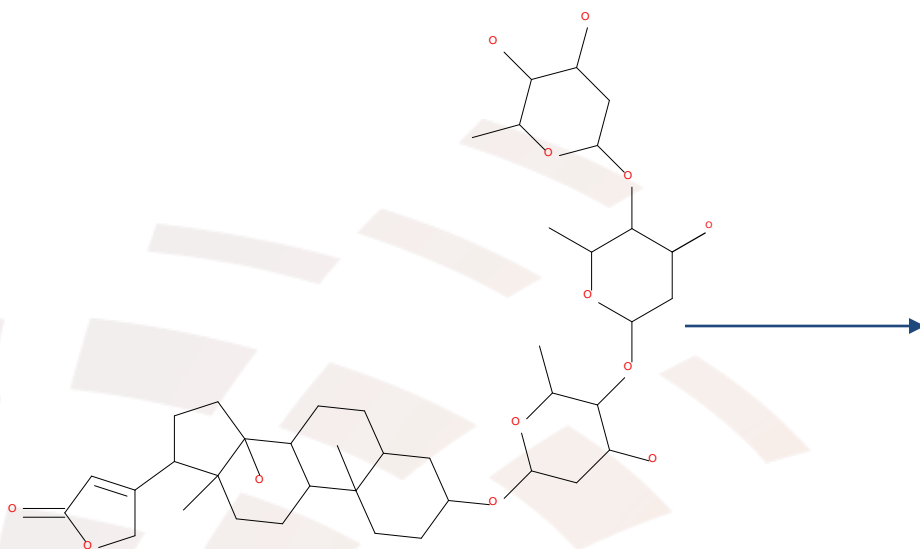


Conformal Prediction in Bioclipse

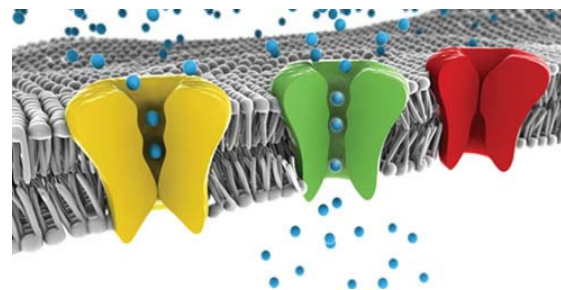
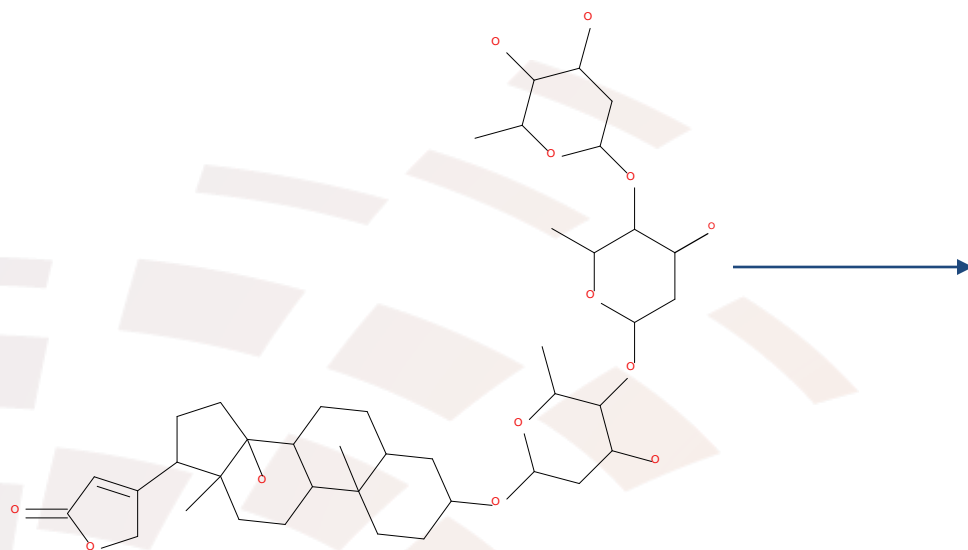
Ernst Ahlberg
Computational ADME & Safety

AstraZeneca 

Computational Safety



Computational Safety

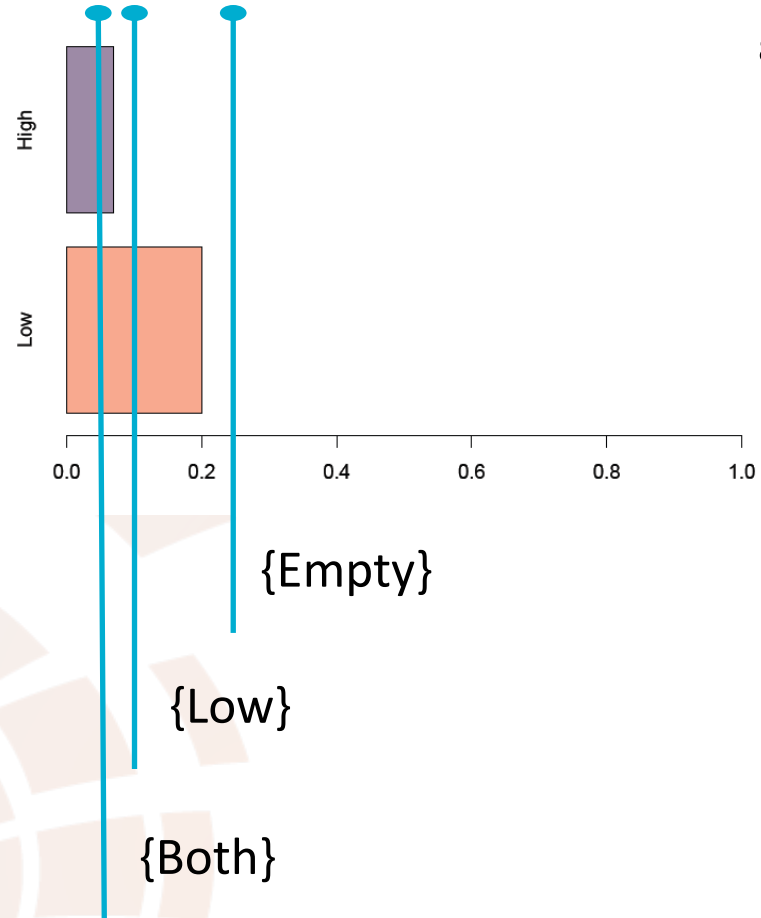


The Conformal Prediction Framework

- **Non Conformity measure:** a measure of how unusual an example looks relative to previous examples, described as a monotonically increasing function.
- **Prediction region:** A set Γ^ε that contains the true value with probability at least $1 - \varepsilon$.
- **Exchangeability:** Consider variables z_1, \dots, z_N . Suppose that for any collection of N values, the $N!$ different orderings are equally likely. Then we say that z_1, \dots, z_N are exchangeable.
- **Validity:** An Γ^ε is valid if it contains the truth $1-\varepsilon$ of the time.
- **Efficiency:** The smallest prediction region at a given ε produces the most informative prediction, ie $\Gamma_1^\varepsilon > \Gamma_2^\varepsilon \Rightarrow \Gamma_2^\varepsilon$ is more efficient than Γ_1^ε . Also, the prediction regions for different ε are nested such that we can obtain a small prediction region with a low level of confidence and with a higher level of confidence the prediction region will be expanded, containing the smaller prediction region. In short when $\varepsilon_1 \geq \varepsilon_2$ we have $\Gamma^{\varepsilon_1} \subseteq \Gamma^{\varepsilon_2}$.

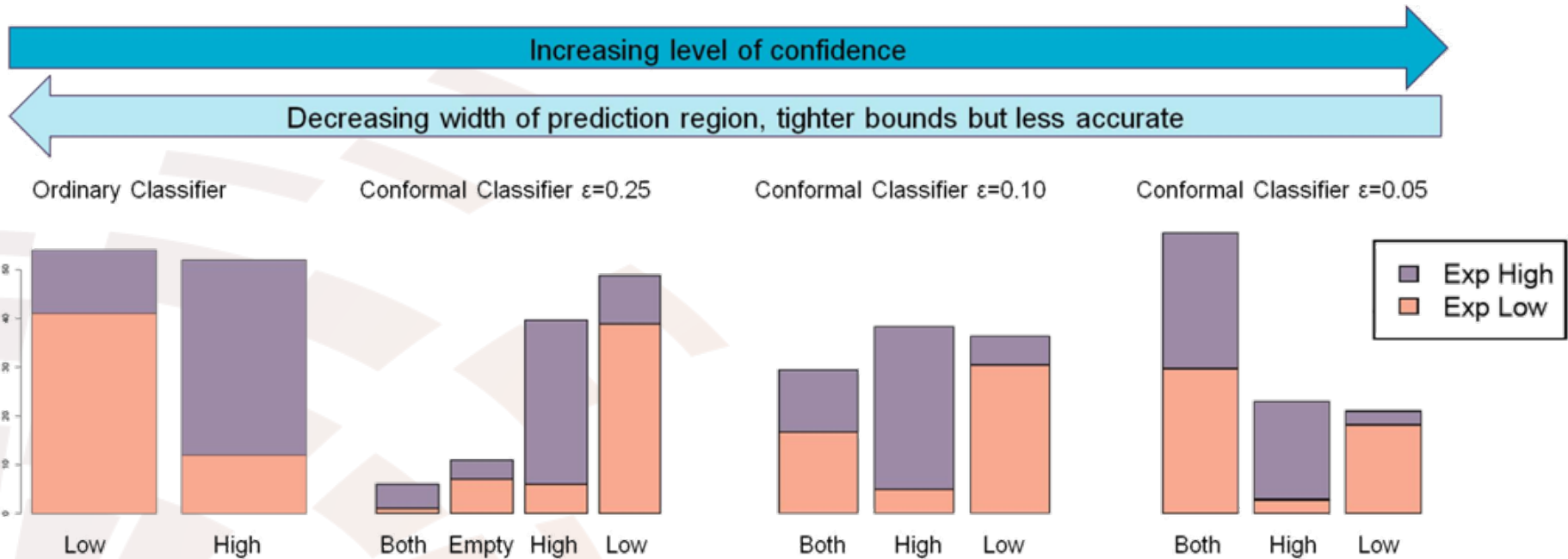
Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005

P-values for a single prediction



Prediction Regions at
 ϵ 0.05, 0.10, 0.25

Compared to an ordinary classifier



Demo

The screenshot displays the Bioclipse software interface. The main window shows the chemical structure of paracetamol (CC(=O)Nc1ccc(O)cc1) with a blue highlight on the hydroxyl group. The interface is divided into several panels:

- Left Panel:** A file browser showing a project named 'demo' containing files like 'danthron.mol', 'paracetamol.mol', and 'Sample Data'.
- Top Panel:** A toolbar with various icons for file operations and a dropdown menu showing 'Default' and 'QSAR'.
- Bottom-Left Panel (Properties):** A table showing dataset and model information.
- Bottom-Right Panel (Decision Support):** A table showing the result of a QSAR model for paracetamol.
- Bottom-Right Panel (Conformal View):** A panel showing the significance level for the prediction.

| Property | Value |
|---------------------|---|
| ▼ Dataset | |
| Dataset name | Bursi Mutagenicity Dataset |
| Descriptors | Signatures (height 0-3) |
| Observations | 4337 |
| URL | http://pubs.acs.org/doi/abs/10.1021/jm0... |
| Variables | 23226 |
| ▼ Model | |
| Learning model | SVM Conformal Prediction |
| Learning parameters | kernel=RBF, c=50, gamma=0.002 |

| Model | Result | Confidence |
|-----------------|------------|----------------|
| ▼ Other | | |
| ▼ Uncategorized | | |
| Ames QSAR | nonmutagen | 0.07 0.34 0.38 |

Conformal View
Significance: 0.14