# *“Compound-kinase binding affinity prediction with confidence guarantees”*

Davor Oršolić, Tomislav Šmuc

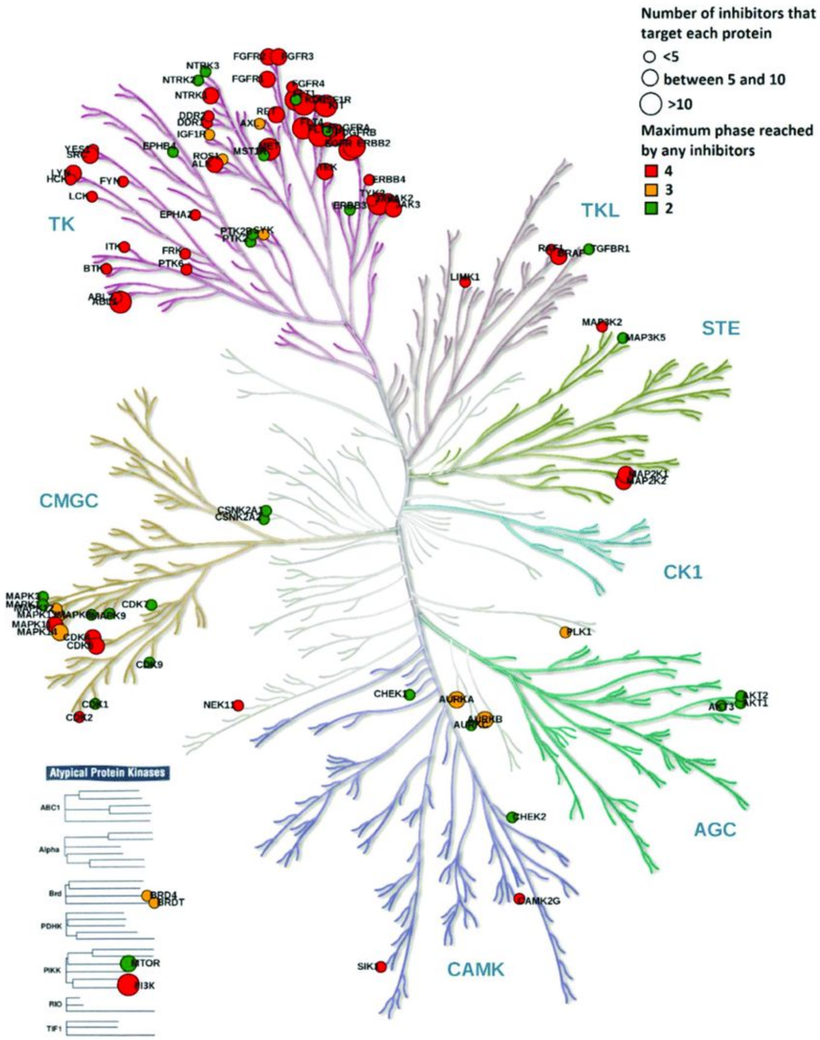
Laboratory for Machine Learning and Knowledge Representation
Ruđer Bošković Institute

# Introduction

- Rapid development in machine learning and computer science allows for efficient profiling of enormous chemical spaces.

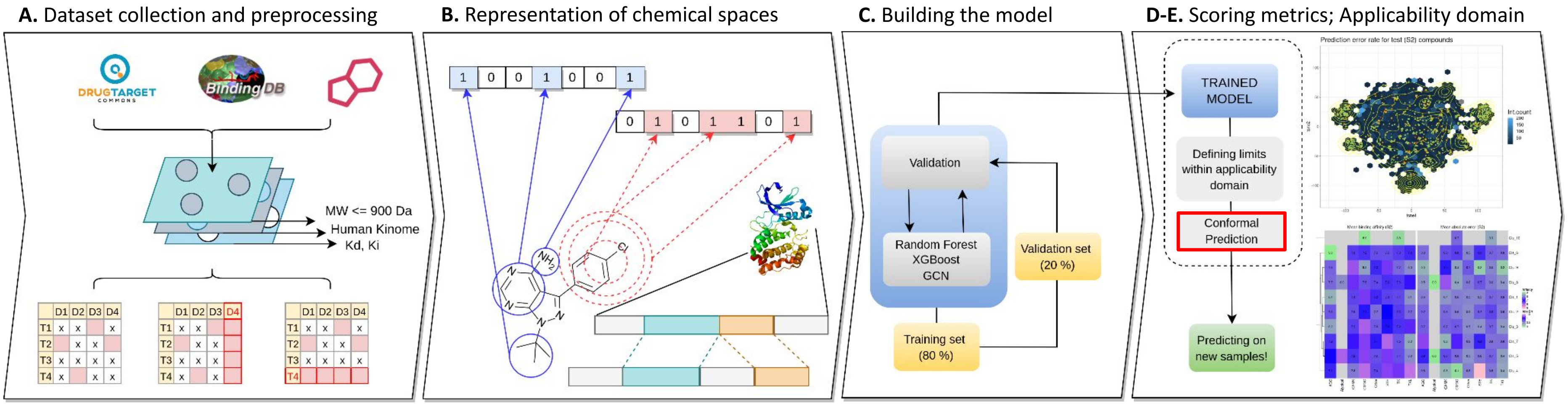
- Protein kinase inhibitors are one of the most popular groups of pharmacologically promising compounds.

- IDG-DREAM Drug-Kinase Binding Prediction Challenge



"Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com)"

# Introduction



**A.** Dataset collection and preprocessing

**B.** Representation of chemical spaces

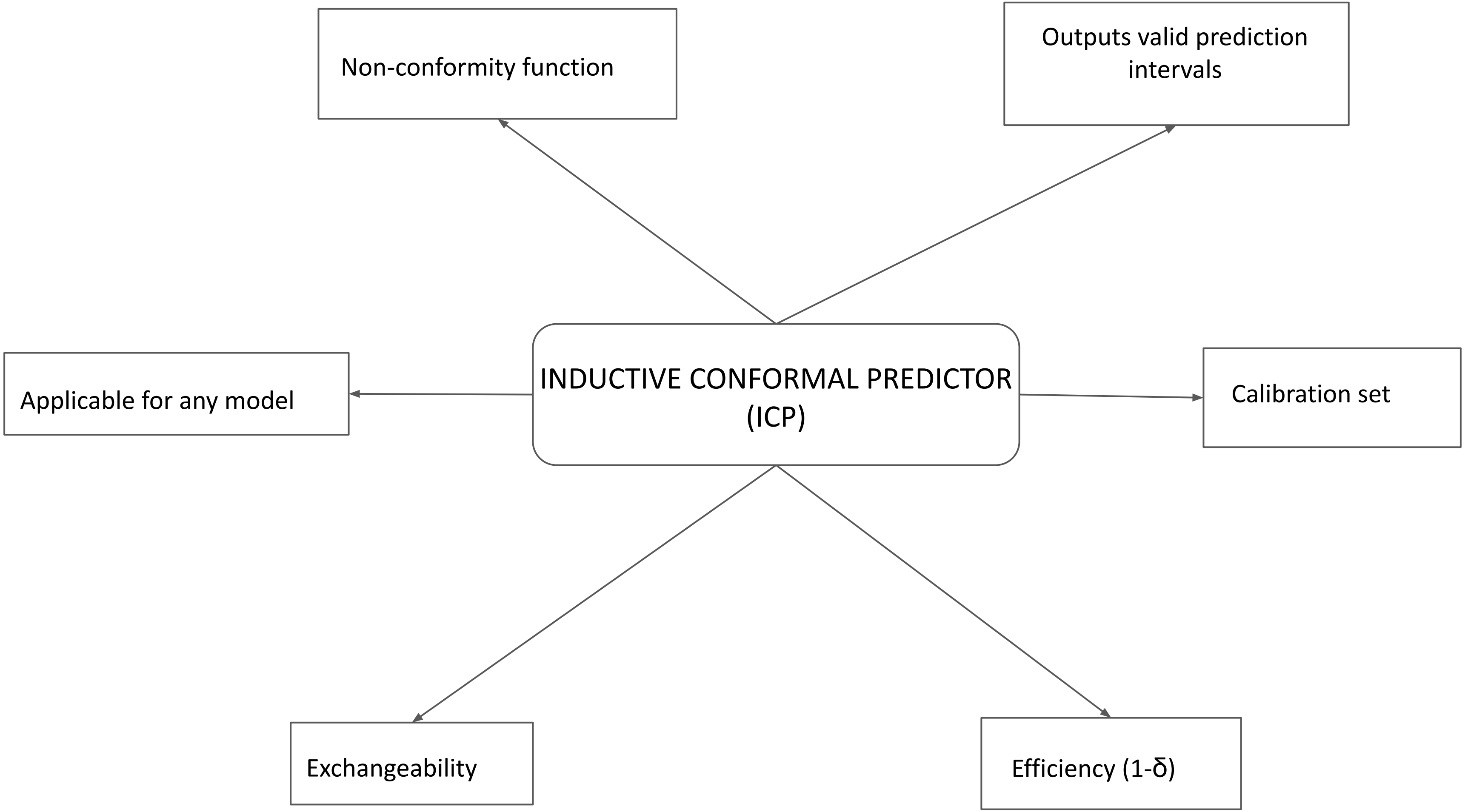
**C.** Building the model

**D-E.** Scoring metrics; Applicability domain

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| T1 | x | x | | x |
| T2 | | x | x | |
| T3 | x | x | x | x |
| T4 | x | | x | x |

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| T1 | x | x | | |
| T2 | | x | x | |
| T3 | x | x | x | |
| T4 | x | | x | |

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| T1 | x | x | | x |
| T2 | | x | x | |
| T3 | x | x | x | x |
| T4 | | | | |



INDUCTIVE CONFORMAL PREDICTOR (ICP)

Non-conformity function

Outputs valid prediction intervals

Applicable for any model

Calibration set

Exchangeability

Efficiency (1-δ)

# Inductive conformal predictor (ICP)

$$Z = \{(x_1,y_1),\ldots,(x_l,y_l)\} \quad \begin{cases} Z^t = \{(x_1,y_1),\ldots,(x_m,y_m)\} \\ Z^c = \{(x_{m+1},y_{m+1}),\ldots,(x_l,y_l)\} \end{cases}$$

For every calibration sample $(x_i,y_i) \in Z^c$

- ➔ Predict output value $\hat{y}_i = h_Z(x_i)$
- ➔ Calculate non-conformity scores $(\alpha_i)$

Non-conformity function:

$$\alpha_i = |y_i - \hat{y}_i|$$

For every tentative label $\tilde{y}$, compute non-conformity score and p-value:

$$p(\tilde{y}) = \frac{\#\left\{z_i \in Z^c \,\middle|\, \alpha_i \geq \alpha_j^{\tilde{y}}\right\} + 1}{|Z^c| + 1}, \; p(\tilde{y}) < \delta$$

Given a significance level $\delta$ and a set of calibration scores $S=\{\alpha_1,\ldots,\alpha_i\}$, locate the smallest $\alpha_{s(\delta)} \in S$ that satisfies the equation:

$$\frac{\#\left\{z_i \in Z^c \,\middle|\, \alpha_i < \alpha_{s(\delta)}\right\} + 1}{|Z^c| + 1} \geq 1 - \delta$$

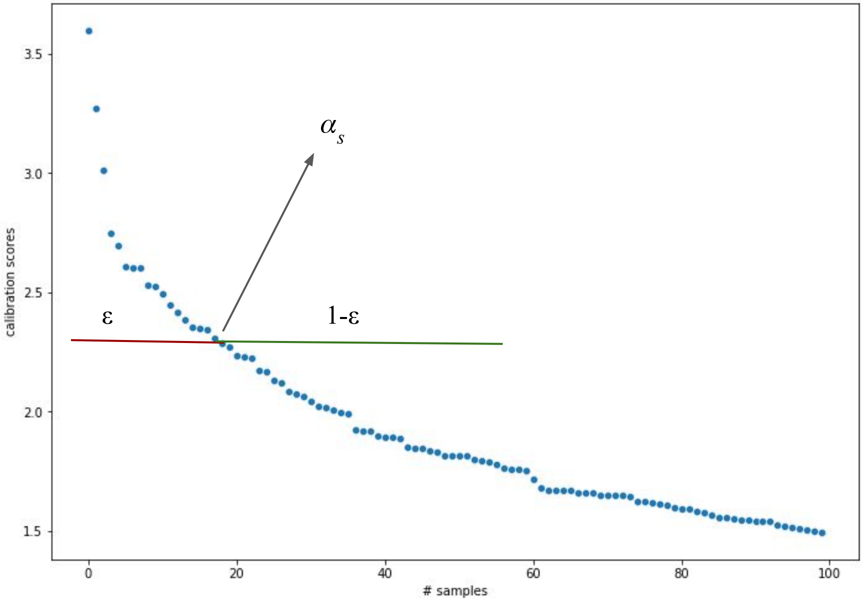
Shafer, G., Vovk, V., 2007. A tutorial on conformal prediction. Journal of Machine Learning Research 9. https://doi.org/10.1145/1390681.1390693

# Inductive conformal predictor (ICP)

$$\frac{\#\{z_i \in Z^c \mid \alpha_i < \alpha_{s(\delta)}\} + 1}{|Z^c| + 1} \geq 1 - \delta$$

$$\Gamma_j^\delta = \hat{\mathrm{y}}_j \pm \alpha_{s(\delta)}$$

$$\Gamma_j^\delta = h_z(x_j) \pm \alpha_{s(\delta)}$$



Compute $\boldsymbol{\alpha}_{\mathbf{x}}$ scores for every tentative $\tilde{\boldsymbol{y}}_{\boldsymbol{x}}$ label?

Johansson, U., Boström, H., Löfström, T., Linusson, H., 2014. Regression conformal prediction with random forests. Machine Learning 97, 1–22. https://doi.org/10.1007/s10994-014-5453-0

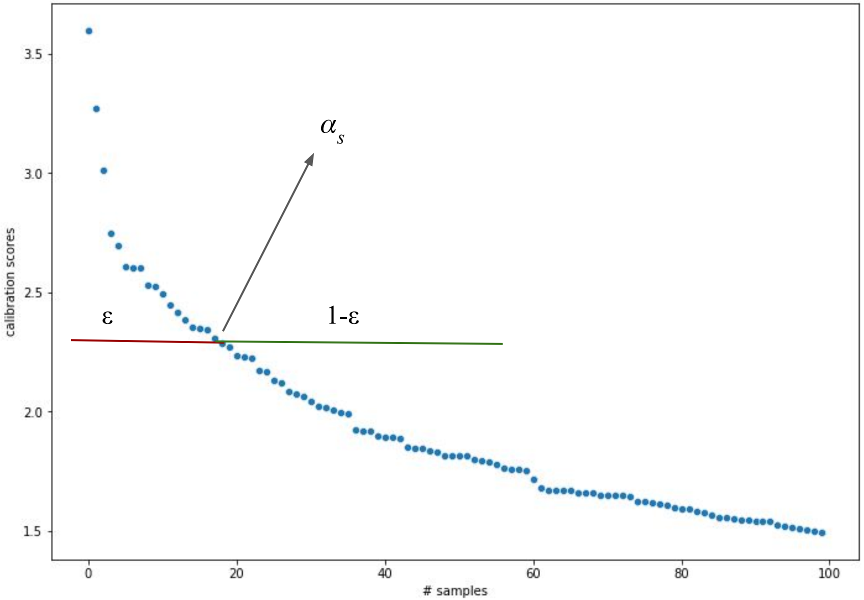
# ICP + Normalisation measure

$$\Gamma_j^\delta = h_z(x_j) \pm \frac{\alpha_{s(\delta)}}{\sigma_j}$$

$$\frac{\#\{z_i \in Z^c \mid \alpha_i < \boxed{\alpha_{s(\delta)}}\} + 1}{|Z^c| + 1} \geq 1 - \delta$$

$$\Gamma_j^\delta = \hat{\mathrm{y}}_j \pm \frac{\alpha_{s(\delta)}}{\sigma_j}$$

Where $\boldsymbol{\sigma_j}$ is an estimate of the accuracy of the underlying model for $\boldsymbol{\hat{y}_j}$.



Other *normalisation* methods include:

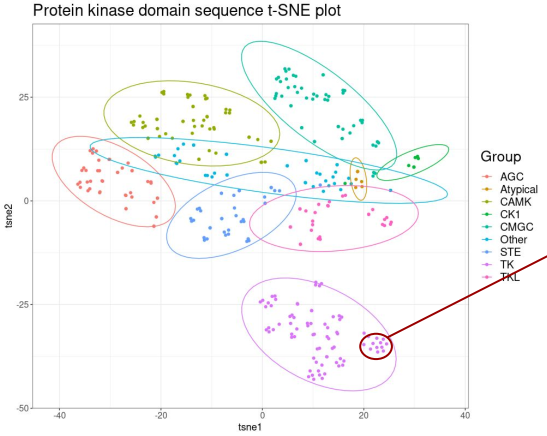
$$\alpha_i = \left|\frac{y_i - \hat{y}_i}{\gamma + \lambda_j^k}\right| \qquad \alpha_i = \left|\frac{y_i - \hat{y}_i}{\gamma + \xi_j^k}\right|$$

Papadopoulos, H., Haralambous, H., 2010. Neural Networks Regression Inductive Conformal Predictor and Its Application to Total Electron Content Prediction, in: Diamantaras, K., Duch, W., Iliadis, L.S. (Eds.), Artificial Neural Networks – ICANN 2010, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 32–41. https://doi.org/10.1007/978-3-642-15819-3_4

Papadopoulos, H., Vovk, V., Gammerman, A., 2011. Regression Conformal Prediction with Nearest Neighbours. jair 40, 815–840. https://doi.org/10.1613/jair.3198

# dAD

Dynamic Applicability Domain

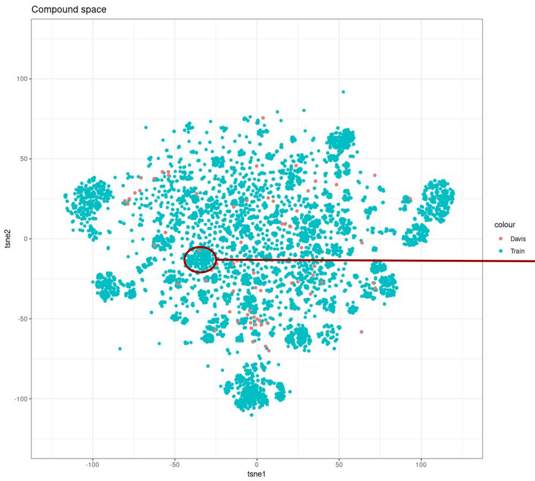

Protein kinase domain sequence t-SNE plot

$$T \subset T^t,\ |T| = q$$

$$\forall t^{(i)} \in T \ \ and \ \ \forall t^{(j)} \in T^t \backslash T$$

$$s\left(t^{(i)}, t\right) \ \geq \ s\left(t^{(j)}, t\right)$$

$$Z^c = \left(x^{(ij)},\ y^{(ij)}\right) : x^{(ij)} \subset (C, T)\ and\ \exists y^{(ij)} \subset Y^t$$


Compound space

$$C \subset C^t,\ |C| = k$$

$$\forall c^{(i)} \in C \ \ and \ \ \forall c^{(j)} \in C^t \backslash C$$

$$s\left(c^{(i)}, c\right) \ \geq \ s\left(c^{(j)}, c\right)$$

# dAD

Dynamic Applicability Domain

$$Z = \{(x_1,y_1),\ldots,(x_l,y_l)\}$$

$$Z^c = \{(x^{(ij)},y^{(ij)}) : x^{(ij)} \subset (C,T) \text{ and } \exists y^{(ij)} \subset Y^t\}$$

$$k=250;\ q=25$$

Where $x^{(ij)}$ represents a tuple $(c^{(i)}, t^{(j)})$.

For every new test sample $x_i$

- ➔ Predict output value $\hat{y}_i = h_Z(x_i);$
- ➔ Locate conformity region in the training space separately for compound (C) and target (T) space;
- ➔ Calculate non-conformity scores $(\alpha_i)$ for calibration samples based on cross-validation predictions (CV) or the sample mean (NN);
- ➔ Calculate non-conformity scores for $x_i$ towards each calibration example, $(\alpha_x)$.

$$\alpha^{cal} = \alpha_i^{nn} = |y_i^{cal} - \bar{y}^{nn}|,\ \alpha^{nn} \in S^{nn}$$

$$\alpha^{cal} = \alpha_i^{cv} = |y_i^{cal} - \hat{y}^{cv}|,\ \alpha^{cv} \in S^{cv}$$

$$\alpha_i^{x} = |y_i^{cal} - \hat{y}|,\ \alpha^{x} \in S^{x}$$

Given a significance level $\delta$ and sets of non-conformity scores for calibration samples $S_i$ and test sample $S_x$, locate the smallest $\alpha_{i(\delta)}$ that satisfies the equation:

$$\frac{\#\{z^{(ij)} \in Z^c \mid \alpha_x < \alpha_{i(\delta)}\} + 1}{|Z^c| + 1} \geq 1 - \delta$$

# Bioactivity space

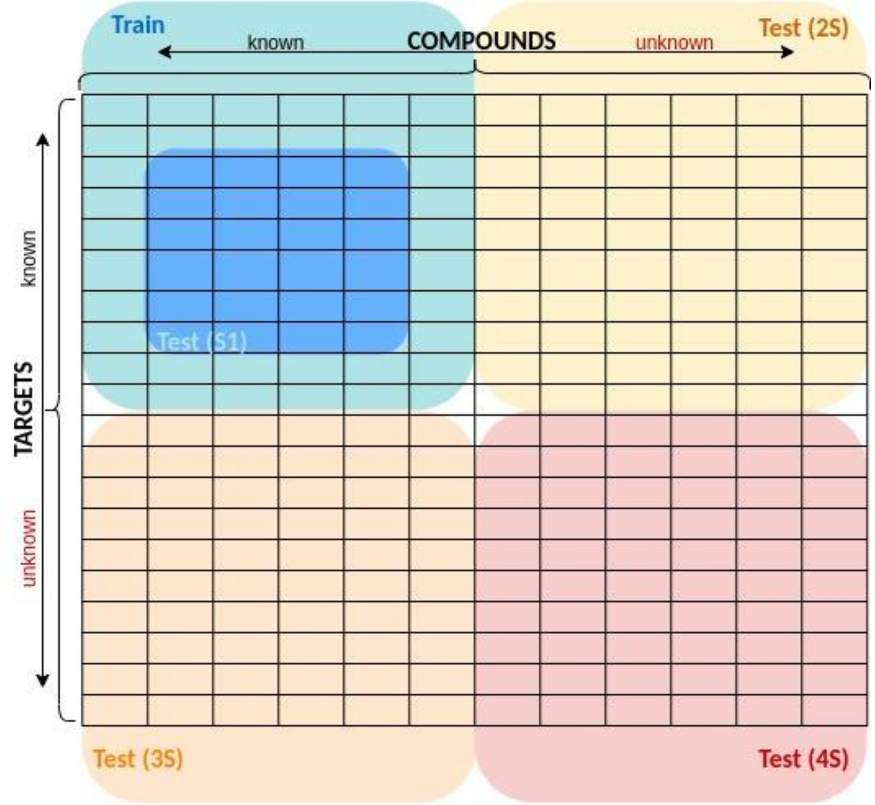
We test this approach on four testing scenarios:

I. contains new compound-target pairs, **S1**;

II. contains new compound-target pairs with compounds never seen in the training set, **S2**;

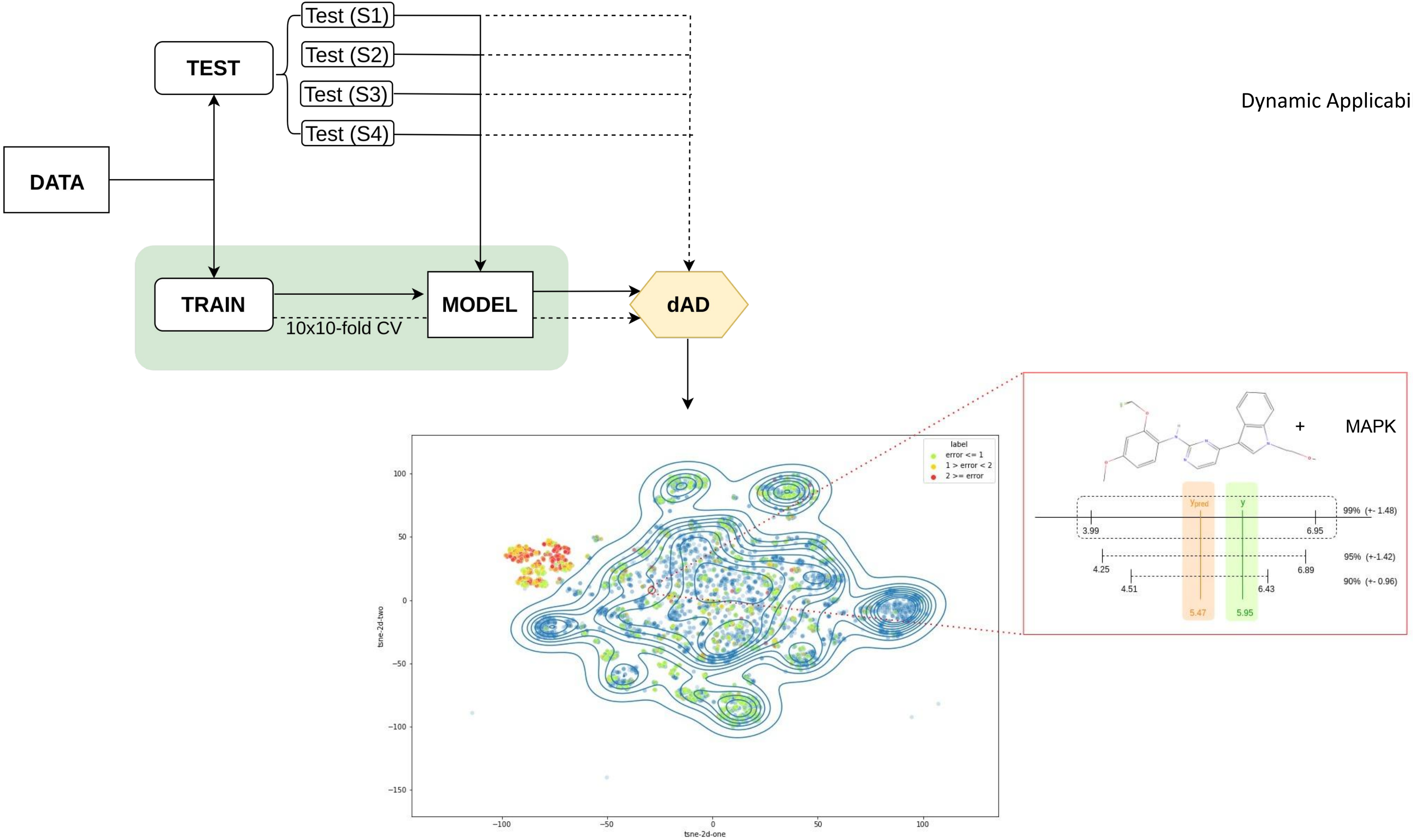
III. contains new compound-target pairs with targets never seen in the training set, **S3**;

IV. contains never seen compounds nor targets in the training set, **S4**.

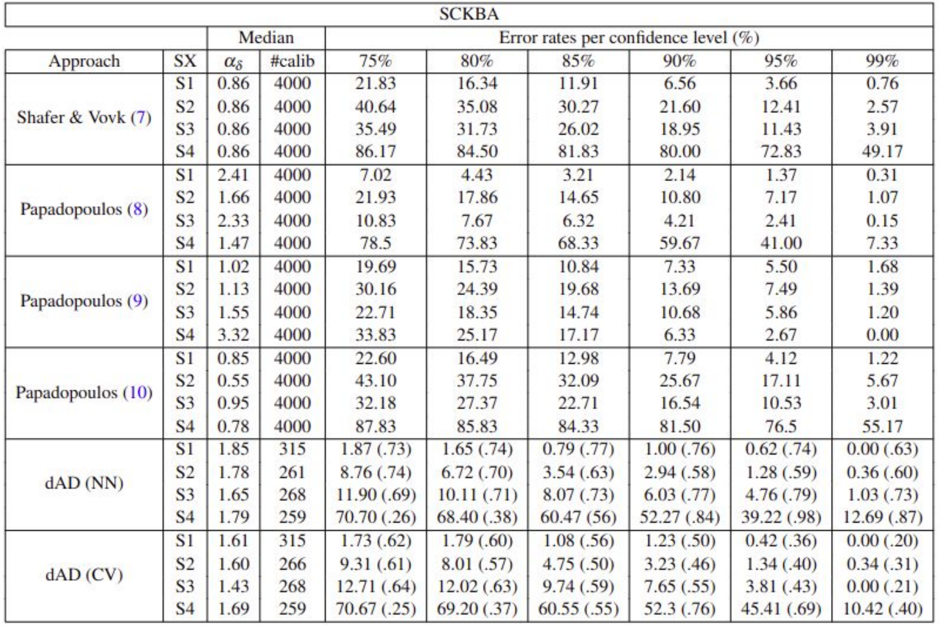


# dAD

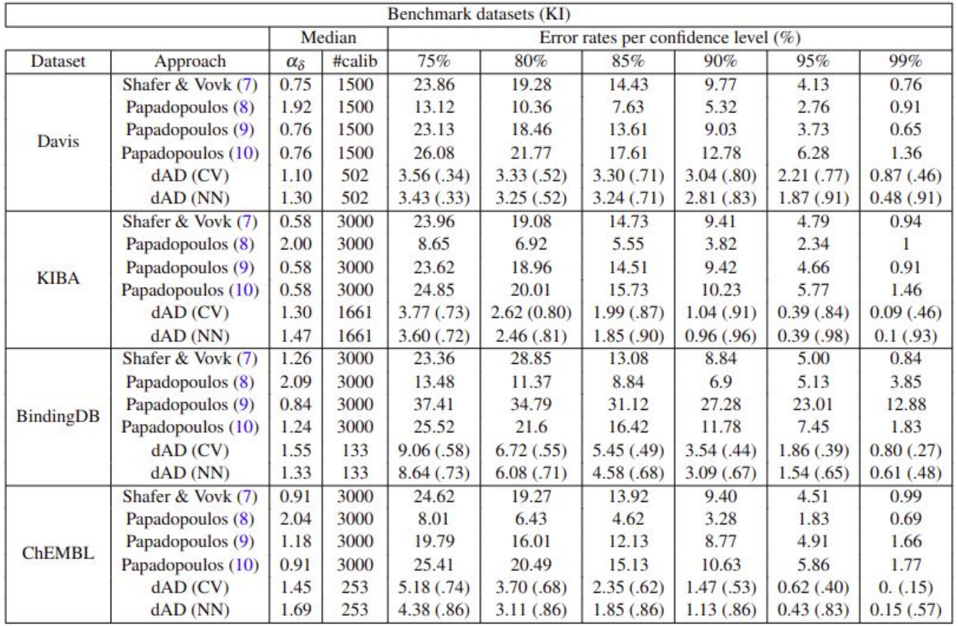
Dynamic Applicability Domain



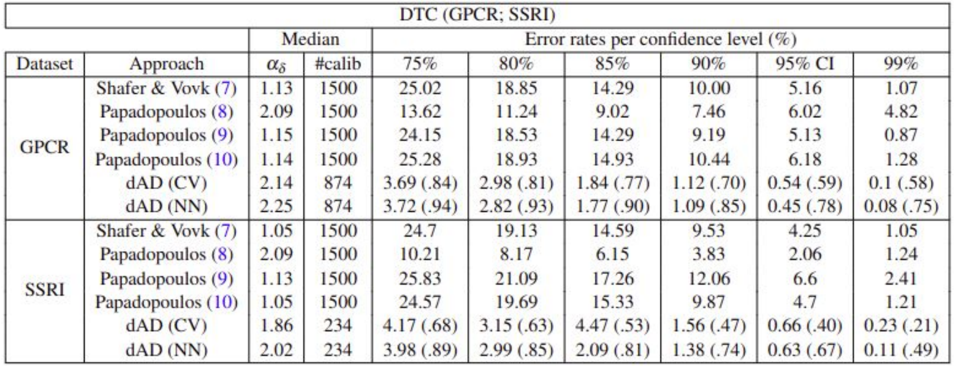
# Baseline comparison



| SCKBA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Median | | Error rates per confidence level (%) | | | | | |
| Approach | SX | $\alpha_\delta$ | #calib | 75% | 80% | 85% | 90% | 95% | 99% |
| Shafer & Vovk (7) | S1 | 0.86 | 4000 | 21.83 | 16.34 | 11.91 | 6.56 | 3.66 | 0.76 |
| | S2 | 0.86 | 4000 | 40.64 | 35.08 | 30.27 | 21.60 | 12.41 | 2.57 |
| | S3 | 0.86 | 4000 | 35.49 | 31.73 | 26.02 | 18.95 | 11.43 | 3.91 |
| | S4 | 0.86 | 4000 | 86.17 | 84.50 | 81.83 | 80.00 | 72.83 | 49.17 |
| Papadopoulos (8) | S1 | 2.41 | 4000 | 7.02 | 4.43 | 3.21 | 2.14 | 1.37 | 0.31 |
| | S2 | 1.66 | 4000 | 21.93 | 17.86 | 14.65 | 10.80 | 7.17 | 1.07 |
| | S3 | 2.33 | 4000 | 10.83 | 7.67 | 6.32 | 4.21 | 2.41 | 0.15 |
| | S4 | 1.47 | 4000 | 78.5 | 73.83 | 68.33 | 59.67 | 41.00 | 7.33 |
| Papadopoulos (9) | S1 | 1.02 | 4000 | 19.69 | 15.73 | 10.84 | 7.33 | 5.50 | 1.68 |
| | S2 | 1.13 | 4000 | 30.16 | 24.39 | 19.68 | 13.69 | 7.49 | 1.39 |
| | S3 | 1.55 | 4000 | 22.71 | 18.35 | 14.74 | 10.68 | 5.86 | 1.20 |
| | S4 | 3.32 | 4000 | 33.83 | 25.17 | 17.17 | 6.33 | 2.67 | 0.00 |
| Papadopoulos (10) | S1 | 0.85 | 4000 | 22.60 | 16.49 | 12.98 | 7.79 | 4.12 | 1.22 |
| | S2 | 0.55 | 4000 | 43.10 | 37.75 | 32.09 | 25.67 | 17.11 | 5.67 |
| | S3 | 0.95 | 4000 | 32.18 | 27.37 | 22.71 | 16.54 | 10.53 | 3.01 |
| | S4 | 0.78 | 4000 | 87.83 | 85.83 | 84.33 | 81.50 | 76.5 | 55.17 |
| dAD (NN) | S1 | 1.85 | 315 | 1.87 (.73) | 1.65 (.74) | 0.79 (.77) | 1.00 (.76) | 0.62 (.74) | 0.00 (.63) |
| | S2 | 1.78 | 261 | 8.76 (.74) | 6.72 (.70) | 3.54 (.63) | 2.94 (.58) | 1.28 (.59) | 0.36 (.60) |
| | S3 | 1.65 | 268 | 11.90 (.69) | 10.11 (.71) | 8.07 (.73) | 6.03 (.77) | 4.76 (.79) | 1.03 (.73) |
| | S4 | 1.79 | 259 | 70.70 (.26) | 68.40 (.38) | 60.47 (56) | 52.27 (.84) | 39.22 (.98) | 12.69 (.87) |
| dAD (CV) | S1 | 1.61 | 315 | 1.73 (.62) | 1.79 (.60) | 1.08 (.56) | 1.23 (.50) | 0.42 (.36) | 0.00 (.20) |
| | S2 | 1.60 | 266 | 9.31 (.61) | 8.01 (.57) | 4.75 (.50) | 3.23 (.46) | 1.34 (.40) | 0.34 (.31) |
| | S3 | 1.43 | 268 | 12.71 (.64) | 12.02 (.63) | 9.74 (.59) | 7.65 (.55) | 3.81 (.43) | 0.00 (.21) |
| | S4 | 1.69 | 259 | 70.67 (.25) | 69.20 (.37) | 60.55 (.55) | 52.3 (.76) | 45.41 (.69) | 10.42 (.40) |

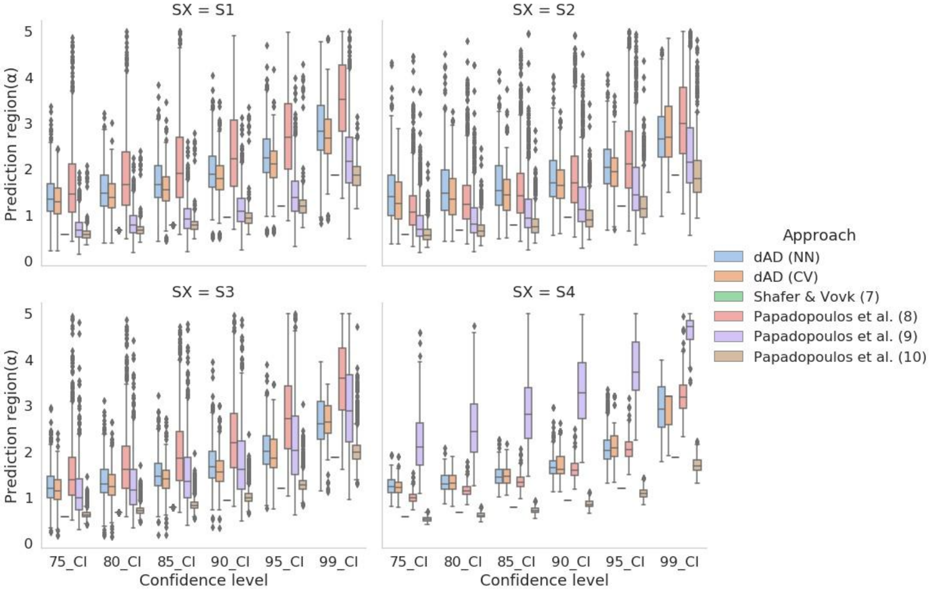


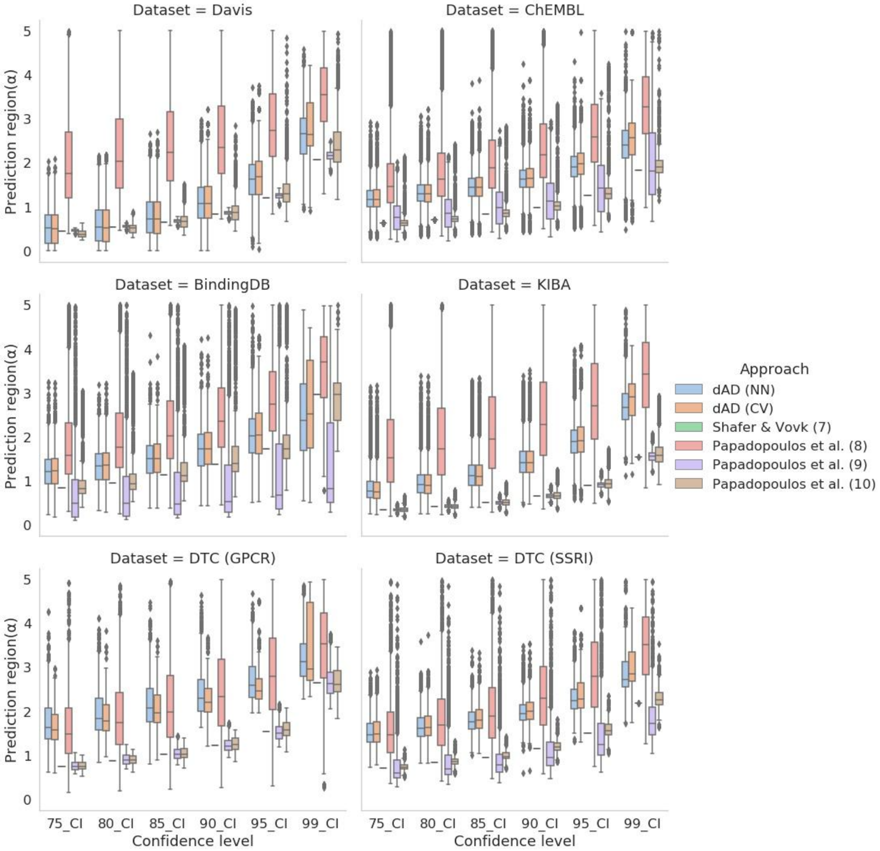
| Benchmark datasets (KI) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Median | | Error rates per confidence level (%) | | | | | |
| Dataset | Approach | $\alpha_\delta$ | #calib | 75% | 80% | 85% | 90% | 95% | 99% |
| Davis | Shafer & Vovk (7) | 0.75 | 1500 | 23.86 | 19.28 | 14.43 | 9.77 | 4.13 | 0.76 |
| | Papadopoulos (8) | 1.92 | 1500 | 13.12 | 10.36 | 7.63 | 5.32 | 2.76 | 0.91 |
| | Papadopoulos (9) | 0.76 | 1500 | 23.13 | 18.46 | 13.61 | 9.03 | 3.73 | 0.65 |
| | Papadopoulos (10) | 0.76 | 1500 | 26.08 | 21.77 | 17.61 | 12.78 | 6.28 | 1.36 |
| | dAD (CV) | 1.10 | 502 | 3.56 (.34) | 3.33 (.52) | 3.30 (.71) | 3.04 (.80) | 2.21 (.77) | 0.87 (.46) |
| | dAD (NN) | 1.30 | 502 | 3.43 (.33) | 3.25 (.52) | 3.24 (.71) | 2.81 (.83) | 1.87 (.91) | 0.48 (.91) |
| KIBA | Shafer & Vovk (7) | 0.58 | 3000 | 23.96 | 19.08 | 14.73 | 9.41 | 4.79 | 0.94 |
| | Papadopoulos (8) | 2.00 | 3000 | 8.65 | 6.92 | 5.55 | 3.82 | 2.34 | 1 |
| | Papadopoulos (9) | 0.58 | 3000 | 23.62 | 18.96 | 14.51 | 9.42 | 4.66 | 0.91 |
| | Papadopoulos (10) | 0.58 | 3000 | 24.85 | 20.01 | 15.73 | 10.23 | 5.77 | 1.46 |
| | dAD (CV) | 1.30 | 1661 | 3.77 (.73) | 2.62 (0.80) | 1.99 (.87) | 1.04 (.91) | 0.39 (.84) | 0.09 (.46) |
| | dAD (NN) | 1.47 | 1661 | 3.60 (.72) | 2.46 (.81) | 1.85 (.90) | 0.96 (.96) | 0.39 (.98) | 0.1 (.93) |
| BindingDB | Shafer & Vovk (7) | 1.26 | 3000 | 23.36 | 28.85 | 13.08 | 8.84 | 5.00 | 0.84 |
| | Papadopoulos (8) | 2.09 | 3000 | 13.48 | 11.37 | 8.84 | 6.9 | 5.13 | 3.85 |
| | Papadopoulos (9) | 0.84 | 3000 | 37.41 | 34.79 | 31.12 | 27.28 | 23.01 | 12.88 |
| | Papadopoulos (10) | 1.24 | 3000 | 25.52 | 21.6 | 16.42 | 11.78 | 7.45 | 1.83 |
| | dAD (CV) | 1.55 | 133 | 9.06 (.58) | 6.72 (.55) | 5.45 (.49) | 3.54 (.44) | 1.86 (.39) | 0.80 (.27) |
| | dAD (NN) | 1.33 | 133 | 8.64 (.73) | 6.08 (.71) | 4.58 (.68) | 3.09 (.67) | 1.54 (.65) | 0.61 (.48) |
| ChEMBL | Shafer & Vovk (7) | 0.91 | 3000 | 24.62 | 19.27 | 13.92 | 9.40 | 4.51 | 0.99 |
| | Papadopoulos (8) | 2.04 | 3000 | 8.01 | 6.43 | 4.62 | 3.28 | 1.83 | 0.69 |
| | Papadopoulos (9) | 1.18 | 3000 | 19.79 | 16.01 | 12.13 | 8.77 | 4.91 | 1.66 |
| | Papadopoulos (10) | 0.91 | 3000 | 25.41 | 20.49 | 15.13 | 10.63 | 5.86 | 1.77 |
| | dAD (CV) | 1.45 | 253 | 5.18 (.74) | 3.70 (.68) | 2.35 (.62) | 1.47 (.53) | 0.62 (.40) | 0. (.15) |
| | dAD (NN) | 1.69 | 253 | 4.38 (.86) | 3.11 (.86) | 1.85 (.86) | 1.13 (.86) | 0.43 (.83) | 0.15 (.57) |



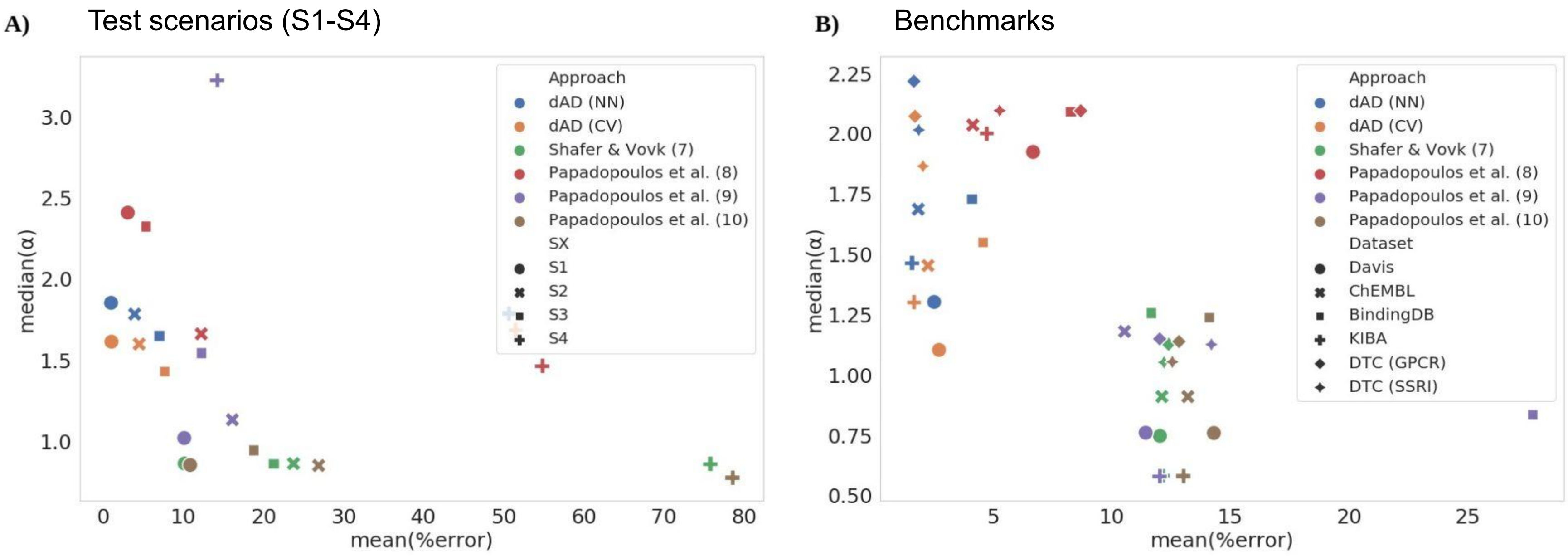
| DTC (GPCR; SSRI) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Median | | Error rates per confidence level (%) | | | | | |
| Dataset | Approach | $\alpha_\delta$ | #calib | 75% | 80% | 85% | 90% | 95% CI | 99% |
| GPCR | Shafer & Vovk (7) | 1.13 | 1500 | 25.02 | 18.85 | 14.29 | 10.00 | 5.16 | 1.07 |
| | Papadopoulos (8) | 2.09 | 1500 | 13.62 | 11.24 | 9.02 | 7.46 | 6.02 | 4.82 |
| | Papadopoulos (9) | 1.15 | 1500 | 24.15 | 18.53 | 14.29 | 9.19 | 5.13 | 0.87 |
| | Papadopoulos (10) | 1.14 | 1500 | 25.28 | 18.93 | 14.93 | 10.44 | 6.18 | 1.28 |
| | dAD (CV) | 2.14 | 874 | 3.69 (.84) | 2.98 (.81) | 1.84 (.77) | 1.12 (.70) | 0.54 (.59) | 0.1 (.58) |
| | dAD (NN) | 2.25 | 874 | 3.72 (.94) | 2.82 (.93) | 1.77 (.90) | 1.09 (.85) | 0.45 (.78) | 0.08 (.75) |
| SSRI | Shafer & Vovk (7) | 1.05 | 1500 | 24.7 | 19.13 | 14.59 | 9.53 | 4.25 | 1.05 |
| | Papadopoulos (8) | 2.09 | 1500 | 10.21 | 8.17 | 6.15 | 3.83 | 2.06 | 1.24 |
| | Papadopoulos (9) | 1.13 | 1500 | 25.83 | 21.09 | 17.26 | 12.06 | 6.6 | 2.41 |
| | Papadopoulos (10) | 1.05 | 1500 | 24.57 | 19.69 | 15.33 | 9.87 | 4.7 | 1.21 |
| | dAD (CV) | 1.86 | 234 | 4.17 (.68) | 3.15 (.63) | 4.47 (.53) | 1.56 (.47) | 0.66 (.40) | 0.23 (.21) |
| | dAD (NN) | 2.02 | 234 | 3.98 (.89) | 2.99 (.85) | 2.09 (.81) | 1.38 (.74) | 0.63 (.67) | 0.11 (.49) |

# Baseline comparison





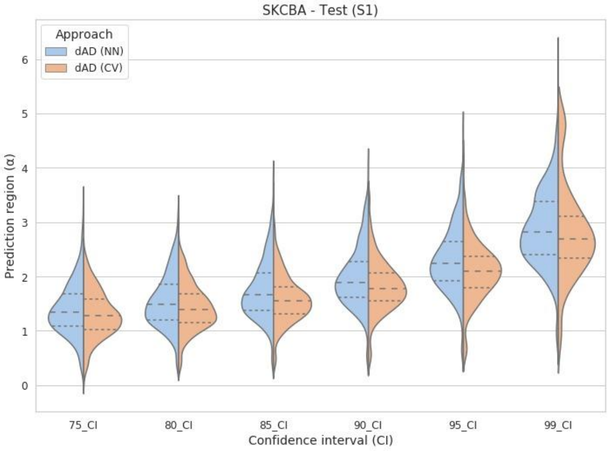
# Baseline comparison

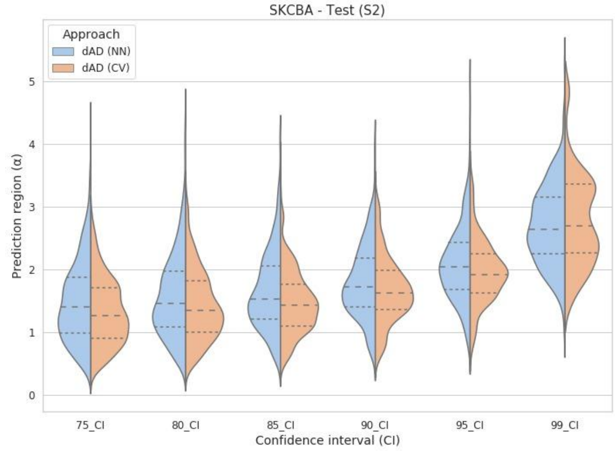


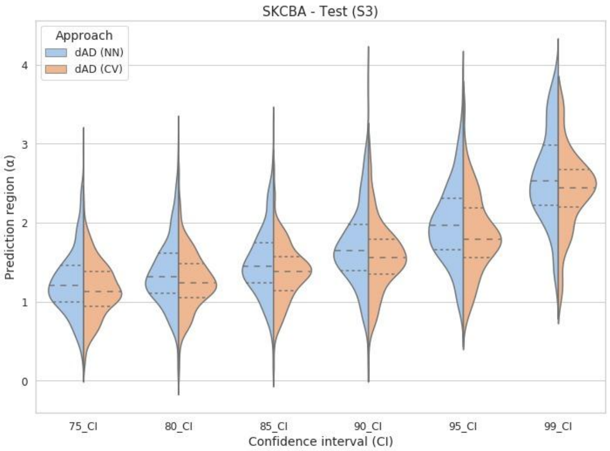
A) Test scenarios (S1-S4)

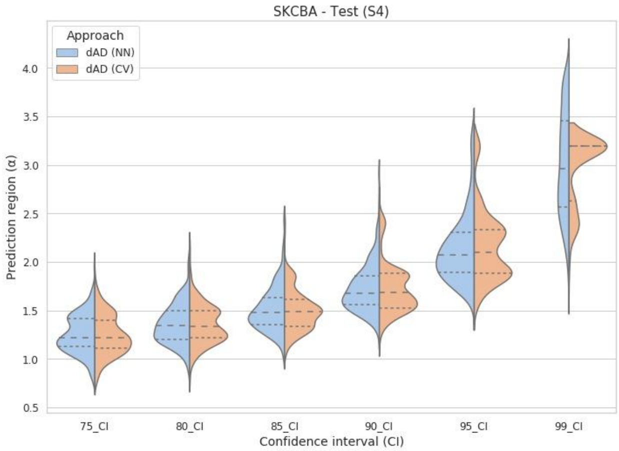
B) Benchmarks

# Baseline comparison









dAD (CV) vs. dAD (NN)

# Concluding remarks

- ➔ **dAD** depends on a sample specific calibration set;
- ➔ Calibration set is defined by the conformity of test compounds and targets individually;
- ➔ Output consists of sample specific prediction regions, with **no** need for additional **normalisation measures**;
- ➔ Provides robust guarantees for suggested prediction regions, and more accurately reflects model performance in the training area close to the test sample;
- ➔ Proved to be more **effective for** challenging prediction settings reflecting **real use-case scenarios** (S2 and S3).

## Acknowledgements:

Research Cooperability Program of the Croatian Science Foundation, funded by the European Union from the European Social Fund under the Operational Programme Efficient Human Resources 2014-2020, through the **Grant 8525**: *Augmented Intelligence Workflows for Prediction, Discovery, and Understanding in Genomics and Pharmacogenomics.*

Hrzz
Croatian Science Foundation

ESF UČINKOVITI LJUDSKI POTENCIJALI

Europska unija
"Zajedno do fondova EU"

EUROPSKI STRUKTURNI I INVESTICIJSKI FONDOVI

ESF UČINKOVITI LJUDSKI POTENCIJALI

Croatian Government and the European Union (European Regional Development Fund — the Competitiveness and Cohesion Operational Program), through the project Bioprospecting of the Adriatic Sea (KK.01.1.1.01.0002), granted to *The Scientific Centre of Excellence for Marine Bioprospecting* — **BioProCro**.

BIOPROSPECTING JADRANSKOG MORA
Projekt sufinancira Europska unija iz Europskog fonda za regionalni razvoj

Republika Hrvatska
Ministarstvo znanosti i obrazovanja

Europska unija
Zajedno do fondova EU

Republika Hrvatska
Ministarstvo regionalnoga razvoja i fondova Europske unije

EUROPSKI STRUKTURNI I INVESTICIJSKI FONDOVI

Operativni program
KONKURENTNOST I KOHEZIJA

Thank you for your attention!