Supporting Information

Quantitative Structure Activity Relationship (QSAR) study predicts small molecule binding to RNA structure

Zhengguo Cai; Martina Zafferani; Olanrewaju M. Akande; Amanda E. Hargrove*

Social Science Research Institute, 140 Science Drive, Durham, NC, 27708, USA
Department of Chemistry, Duke University, 124 Science Drive, Durham, NC 27708, USA

*Corresponding author: amanda.hargrove@duke.edu

Table of Contents

| Section | Description | Page |
|---------|-------------|------|
| A       | Supplementary table and figures | S-2  |
|         | Table S1. Parameters used for conformation search | S-2  |
|         | Figure S1. Natural log transformation of response variables | S-2  |
|         | Figure S2. Lasso selection of $\ln k_{on}/\ln k_{off}$ descriptors and baseline models | S-3  |
|         | Figure S3. Q-Q plot and Williams plot for $\ln k_{on}/\ln k_{off}$ models | S-4  |
|         | Figure S4. Plots of fitting residuals against each descriptor from 3 MLR models | S-5  |
|         | Figure S5. Plots of fitting residuals against fitted values for 3 MLR models | S-5  |
|         | Figure S6. Model stability test of $\ln k_{on}/\ln k_{off}$ models | S-6  |
| B       | Chemistry | S-7  |
|         | Chemical structures of molecules for model training | S-7  |
|         | Synthesis and characterization of diminazenes | S-8  |
| C       | Surface plasmon resonance | S-18 |
|         | Sensorgrams and fitting parameters | S-18 |
| D       | QSAR modeling | S-116 |
|         | Descriptor calculation | S-116 |
|         | Methods and scripts | S-117 |
|         | Descriptor refinement | S-117 |
|         | Representative data splitting by Kennard-Stone algorithm and PCA | S-119 |
|         | Descriptor selection by lasso and model selection | S-121 |
|         | Ensemble learning-based models | S-124 |
|         | Model assessment: Q-Q plot and Williams plot | S-128 |
|         | Predictor stability test | S-130 |
| E       | References | S-132 |
Section A. Supplementary table and figures

| Parameter                  | Input  |
|----------------------------|--------|
| Rejection limit            | 100    |
| Iteration limit            | 10000  |
| RMS gradient               | 0.005  |
| MM iteration limit         | 500    |
| RMSD limit                 | 0.15   |
| Energy window              | 3      |
| Conformation limit         | 10000  |

Table S1. Parameters used for conformation search

Figure S1 Natural log transformation was taken for each response variable to shift the skewed distribution close to a normal distribution.
A. Lasso selection of $\ln k_{on}$ descriptors

Number of non-zero descriptors

Coeficients

Log ($\lambda$)

B. Baseline model of $\ln k_{on}$

$\ln k_{on} = -12.27\text{GCUT}_\text{PE0E}_e + 0.093\text{vs}-\text{other} + 0.59\text{vsurf}_\text{DW12} + 0.42\text{vsurf}_\text{DD23}$

$R^2_{training} = 0.77$

$Q^2_{test} = 0.77$

C. Lasso selection of $\ln k_{off}$ descriptors

Number of non-zero descriptors

Coeficients

Log ($\lambda$)

D. Baseline model of $\ln k_{off}$

$\ln k_{off} = 2.0 - 0.69a_{base} - 0.42a_{nN} + 0.27\text{vsurf}_\text{DD13}$

$R^2_{training} = 0.64$

$Q^2_{test} = 0.61$

Figure S2 A. Lasso selection of $\ln k_{on}$ descriptors, the best $\lambda$ was determined as 0.22 from 5-fold cross validation. B. Observed $\ln k_{on}$ was plotted with the value predicted by the MLR baseline model shown at top. C. Lasso selection of $\ln k_{off}$ descriptors, the optimized $\lambda$ was determined as 0.14 to ensure the inclusion of a decisive descriptor: vsurf_DD13. D. Observed $\ln k_{off}$ was plotted with the value predicted by the MLR baseline model shown at top.
Figure S3 A. Normal quantile-quantile plots of ln$k_{on}$ and ln$k_{off}$ models. B. Williams plot showed applicable domain of ln$k_{on}$ and ln$k_{off}$ models with training and test sets.
Figure S4 Plots of fitting residuals against each descriptor from 3 MLR models (A. $\ln K_0$ model. B. $\ln k_{on}$ model. C. $\ln k_{off}$ model) to check linearity assumption.

Figure S5 Plots of fitting residuals against the fitted values for 3 MLR models to check independence and equal variance assumption.
A. Train/test stability of ln\textsubscript{on} model

\[ \text{ln}\textsubscript{on} \sim 1 + \text{GCUT\_PEOE}\_0 + \text{vsa\_other} + \text{vsurf}\_DW12 + \text{vsurf}\_DD23 \]

\[ \text{R}^2\text{\_training} = 0.78 \pm 0.03 \]

\[ \text{Q}^2\text{\_test} = 0.69 \pm 0.15 \]

B. Train/test stability of ln\textsubscript{off} model

\[ \text{ln}\textsubscript{off} \sim 1 + \text{a\_base} + \text{a\_nN} + \text{vsurf}\_DD13 \]

\[ \text{R}^2\text{\_training} = 0.65 \pm 0.04 \]

\[ \text{Q}^2\text{\_test} = 0.56 \pm 0.17 \]

Figure S6 A. Model stability test on ln\textsubscript{on} data using formula: ln\textsubscript{on} \sim 1 + GCUT\_PEOE\_0 + vsa\_other + vsurf\_DW12 + vsurf\_DD23. B. Model stability test on ln\textsubscript{off} data using formula: ln\textsubscript{off} \sim 1 + a\_base + a\_nN + vsurf\_DD13.
Section B. Chemistry

1. Chemical structures of molecules for model training

Figure S7. Chemical structures of molecules for model training: DMA-1~DMA-164 are from ref 1, DMA-180~DMA-194 from ref 2, DMA compounds from ref 3, DPF x1~DPF x10 from ref 4 (x = m or p), DPF p13, p15 from ref 5. DMZs were synthesized as below. The rest of compounds are commercially available.
2. Synthesis and characterization of diminazenes (DMZ)

Reaction schemes and DMZ structures

Scheme S1. Synthetic routes for DMZ compounds

Figure S8. Chemical structures of DMZ synthetic intermediates and three DMZs used in this work
Characterization spectra

- DMZ M3

**Figure S9 A.** The $^1$H-NMR spectrum of DMZ m3
**Figure S9 B.** The $^{13}$C-NMR spectrum of DMZ m3
Figure S9 C. The HPLC spectrum of DMZ m3
Figure S10 A. The $^1$H-NMR spectrum of DMZ p8

- DMZ P8
Figure S10 B. The $^{13}$C-NMR spectrum of DMZ p8
Figure S10 C. The HPLC spectrum of DMZ p8
Figure S11 A. The $^1$H-NMR spectrum of DMZ p13
**Figure S11 B.** The $^{13}$C-NMR spectrum of DMZ p13
**<Sample Information>**

Sample Name: DMZ-P13  
Sample ID: DMZ-P13  
Data Filename: DMZ-P13.lcd  
Method Filename: GP short-Grd10-90_22min_PDA.lcm  
Batch Filename: 10282020_MSHECK.lcd  
Vial #: 1-17  
Injection Volume: 10 uL  
Date Acquired: 10/29/2020 1:29:27 AM  
Date Processed: 10/29/2020 1:51:30 AM  
Sample Type: Unknown  
Acquired by: chemist  
Processed by: chemist

**<Chromatogram>**

![HPLC Spectrum](image)

**<Peak Table>**

| Peak# | Ret. Time | Area% |
|-------|-----------|-------|
| Total |           |       |

| PDA Ch1 254nm | Ret. Time | Area% |
|---------------|-----------|-------|
| 1             | 10.409    | 1.689 |
| 2             | 11.565    | 98.311|
| Total         |           | 100.000|

**Figure S11 C.** The HPLC spectrum of DMZ p13
Section C. Surface plasmon resonance

Sensorgrams, fitting parameters and quality control table

The units for $k_{on}$, $k_{off}$ and $K_D$ are M$^{-1}$s$^{-1}$, s$^{-1}$, and M, respectively.

- Neomycin

| Curve          | ka (1/Ms)     | kd (1/s)     | KD (M)     | Rmax (RU) | Conc (M) |
|----------------|---------------|--------------|------------|-----------|----------|
| Cycle: 39      | 0.03125 µM    | 2.4E+5       | 0.03125    | 8.5E-8    | 41.22    |
| Cycle: 40      | 0.0625 µM     | 6.35E-8      | 0.0625     | 3.125E-8  | 41.22    |
| Cycle: 41      | 0.125 µM      | 1.25E-7      | 0.125      | 3.125E-8  | 41.22    |
| Cycle: 42      | 0.25 µM       | 2.5E-7       | 0.25       | 3.125E-8  | 41.22    |
| Cycle: 43      | 0.5 µM        | 5.0E-7       | 0.5        | 3.125E-8  | 41.22    |
| Cycle: 44      | 0.03125 µM    | 3.125E-8     | 0.03125    | 8.5E-8    | 41.22    |

Quality Control Report Parameters:

- Kinetic constants are within instrument specifications.
- Kinetic constants appear to be uniquely determined.
- No significant bulk contributions (B) found.
- Check that sensorgrams have sufficient contrast.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S12. The SPR sensorgrams, fitting parameters and quality control table of neomycin (3 replicates)
- Paromomycin

### Curve Data

| Curve   | $ka$ (1/Ms) | $kd$ (1/s) | $KD$ (M) | $R_{max}$ (RU) | Conc (M) |
|---------|-------------|------------|----------|----------------|----------|
| Cycle: 41 3.75 µM | 1.160E+5 | 0.1787 | 1.540E-6 | 3.750E-6 | 65.55 |
| Cycle: 42 7.5 µM | | | | 7.500E-6 |
| Cycle: 43 15 µM | | | | 1.500E-5 |
| Cycle: 44 30 µM | | | | 3.000E-5 |
| Cycle: 45 15 µM | | | | 1.500E-5 |

**Quality Control**
- Kinetic constants are within instrument specifications.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (RU) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plots. Pay attention to systematic and non-random deviations.

### Curve Data

| Curve   | $ka$ (1/Ms) | $kd$ (1/s) | $KD$ (M) | $R_{max}$ (RU) | Conc (M) |
|---------|-------------|------------|----------|----------------|----------|
| Cycle: 9 1.875 µM | 6.335E+4 | 0.3724 | 5.833E-6 | 83.83 | 1.875E-6 |
| Cycle: 10 3.75 µM | | | | 3.750E-6 |
| Cycle: 11 7.5 µM | | | | 7.500E-6 |
| Cycle: 12 15 µM | | | | 1.500E-5 |
| Cycle: 13 30 µM | | | | 3.000E-5 |
| Cycle: 14 7.5 µM | | | | 7.500E-6 |
Figure S13. The SPR sensorgrams, fitting parameters and quality control table of paromomycin (3 replicates)
- Sisomycin

| Curve   | ka (1/Hs) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|---------|-----------|----------|--------|-----------|----------|
| Cycle: 16 1.875 μM | 5.869E+4  | 0.3859  | 6.576E-6 | 77.62     | 1.875E-6 |
| Cycle: 17 3.75 μM |           |          |        |           | 3.750E-6 |
| Cycle: 18 7.5 μM |           |          |        |           | 7.500E-6 |
| Cycle: 19 15 μM |           |          |        |           | 1.500E-5 |
| Cycle: 20 30 μM |           |          |        |           | 3.000E-5 |
| Cycle: 21 7.5 μM |           |          |        |           | 7.500E-6 |

Quality Control Report Residual Parameters

- Kinetic constant kd is approaching the limits that can be measured by the instrument.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (RI) found.
- Check that sensograms have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S14. The SPR sensorgrams, fitting parameters and quality control table of sisomycin (3 replicates)
- Streptomycin

**Graph 1**

| Curve   | $k_a$ (1/Ms) | $k_d$ (1/s) | $K_D$ (M) | $R_{max}$ (RU) | $C_{conc}$ (M) |
|---------|--------------|-------------|-----------|----------------|----------------|
| Cycle: 23 1.875 µM | 845.9        | 0.009334    | 1.163E-5  | 27.48          | 1.875E-5       |
| Cycle: 24 3.75 µM  |              |             |           |                | 3.750E-6       |
| Cycle: 25 7.5 µM   |              |             |           |                | 7.500E-5       |
| Cycle: 26 15 µM    |              |             |           |                | 1.500E-5       |
| Cycle: 27 30 µM    |              |             |           |                | 3.000E-5       |
| Cycle: 28 7.5 µM   |              |             |           |                | 7.500E-5       |

**Graph 2**

| Curve | $k_e$ (1/Ms) | $k_d$ (1/s) | $K_D$ (M) | $R_{max}$ (RU) | $C_{conc}$ (M) |
|-------|--------------|-------------|-----------|----------------|----------------|
| Cycle: 35 3.75 µM | 153.4        | 0.001721    | 1.086E-5  | 22.30          | 3.750E-6       |
| Cycle: 36 7.5 µM  |              |             |           |                | 7.500E-6       |
| Cycle: 37 15 µM   |              |             |           |                | 1.500E-5       |
| Cycle: 38 30 µM   |              |             |           |                | 3.000E-5       |
| Cycle: 39 15 µM   |              |             |           |                | 1.500E-5       |

**Notes:**

- Kinetic constant $k_a$ is approaching the limits that can be measured by the instrument.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (R) found.
- Check that sensors have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S15. The SPR sensorgrams, fitting parameters and quality control table of streptomycin (3 replicates)
- Tobramycin

![Graph 1](image1)

| Curve  | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|--------|-----------|----------|--------|-----------|----------|
| Cycle: 30  1.875 µM | 5.855E+4  | 0.3347  | 5.717E-6 | 74.82    |          |
| Cycle: 31  3.75 µM  |          |          |        |           | 1.875E-5 |
| Cycle: 32  7.5 µM   |          |          |        |           | 3.750E-5 |
| Cycle: 33  15 µM    |          |          |        |           | 7.500E-5 |
| Cycle: 34  30 µM    |          |          |        |           | 1.500E-5 |
| Cycle: 35  7.5 µM   |          |          |        |           | 3.000E-5 |

![Graph 2](image2)

| Curve  | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|--------|-----------|----------|--------|-----------|----------|
| Cycle: 23  1.875 µM | 5.772E+4  | 0.1532  | 2.654E-6 | 63.32    |          |
| Cycle: 24  3.75 µM  |          |          |        |           | 1.875E-6 |
| Cycle: 25  7.5 µM   |          |          |        |           | 3.750E-6 |
| Cycle: 26  15 µM    |          |          |        |           | 7.500E-6 |
| Cycle: 27  30 µM    |          |          |        |           | 1.500E-5 |
| Cycle: 28  7.5 µM   |          |          |        |           | 3.000E-5 |
Figure S16. The SPR sensorgrams, fitting parameters and quality control table of tobramycin (3 replicates)
• Gentamycin

**Graph and Table: Gentamycin Data**

| Cycle | Concentration (M) | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) |
|-------|------------------|-----------|----------|--------|-----------|
| 2     | 1.875 µM         | 3.899E+4  | 6.231    | 3.235E-6 | 65.40     |
| 3     | 3.75 µM          |           |          |        |           |
| 4     | 7.5 µM           |           |          |        |           |
| 5     | 15 µM            |           |          |        |           |
| 6     | 30 µM            |           |          |        |           |
| 7     | 7.5 µM           |           |          |        |           |

**Graph and Table: Gentamycin Data**

| Cycle | Concentration (M) | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) |
|-------|------------------|-----------|----------|--------|-----------|
| 37    | 1.875 µM         | 6.148E+4  | 0.2648   | 4.307E-6 | 109.5     |
| 38    | 3.75 µM          |           |          |        |           |
| 39    | 7.5 µM           |           |          |        |           |
| 40    | 15 µM            |           |          |        |           |
| 41    | 30 µM            |           |          |        |           |
| 42    | 7.5 µM           |           |          |        |           |

**Quality Control Report**
- Residuals: Parameters
- Check that samplegrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S17. The SPR sensorgrams, fitting parameters and quality control table of gentamycin (3 replicates)
• Neamine
Figure S18. The SPR sensorgrams, fitting parameters and quality control table of neamine (3 replicates)
• Kanamycin

| Curve   | ka (1/Ms) | kd (1/s) | KD (M)   | Rmax (RU) | Conc (M) |
|---------|-----------|----------|----------|-----------|----------|
| Cycle: 16 1.875 µM | 5.029E+4   | 1.015    | 2.018E-5 | 75.66     | 1.675E-6 |
| Cycle: 17 3.75 µM    |           |          |          |           | 3.750E-6 |
| Cycle: 18 7.5 µM     |           |          |          |           | 7.500E-6 |
| Cycle: 19 15 µM      |           |          |          |           | 1.500E-5 |
| Cycle: 20 30 µM      |           |          |          |           | 3.000E-5 |
| Cycle: 21 7.5 µM     |           |          |          |           | 7.500E-6 |

- Kinetic constant kd is approaching the limits that can be measured by the instrument.
- Kinetic constants appear to be unusually determined.
- High bulk contributions (RI) found.
- Check that sensors have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and nonrandom deviations.
Figure S19. The SPR sensorgrams, fitting parameters and quality control table of kanamycin (3 replicates)
- Amikacin

| Curve | ka (1/Hs) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|-------|-----------|----------|--------|-----------|----------|
| Cycle: 5 3.75 μM | 7.498E+4 | 0.5620 | 7.76E-6 | 30.65 | 3.75E-6 |
| Cycle: 6 7.5 μM | | | | | 7.50E-6 |
| Cycle: 7 15 μM | | | | | 1.50E-5 |
| Cycle: 8 30 μM | | | | | 3.00E-5 |
| Cycle: 9 15 μM | | | | | 1.50E-5 |

Quality Control | Report | Residuals | Parameters
--- | --- | --- | ---
• Athentic constant kd is approaching the limits that can be measured by the instrument.
• Kinetic constants appear to be uniquely determined.
• High bulk contributions (R) found.
• Check that sensorgrams have sufficient curvature.
• Examine the residual plot. Pay attention to systematic and non-random deviations.

| Curve | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|-------|-----------|----------|--------|-----------|----------|
| Cycle: 23 1.875 μM | 3.987E+4 | 1.025 | 2.57E-5 | 97.03 | 1.675E-6 |
| Cycle: 24 3.75 μM | | | | | 3.75E-6 |
| Cycle: 25 7.5 μM | | | | | 7.50E-6 |
| Cycle: 26 15 μM | | | | | 1.50E-5 |
| Cycle: 27 30 μM | | | | | 3.00E-5 |
| Cycle: 28 7.5 μM | | | | | 7.50E-6 |
Figure S20. The SPR sensorgrams, fitting parameters and quality control table of amikacin (3 replicates)
• DMA-1

### Table

| Curve   | ka (1/μs) | kd (1/s) | KD (M)  | Rmax (RU) | Conc (M)  |
|---------|-----------|----------|---------|-----------|-----------|
| Cycle: 23 25 μM | 2.10E+4 | 2.41E-1 | 1.15E-4 | 20.46     | 2.50E-5   |
| Cycle: 24 37.5 μM |          |          |         |           |           |
| Cycle: 25 50 μM |          |          |         |           |           |
| Cycle: 26 75 μM |          |          |         |           |           |
| Cycle: 27 75 μM |          |          |         |           |           |

### Notes
- Kinetic constant kd is outside the limits that can be measured by the instrument.
- Kinetic constants cannot be uniquely determined.
- Bulk contributions (R) were not evaluated: The R parameter is set to constant.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S21. The SPR sensorgrams, fitting parameters and quality control table of DMA-1 (3 replicates)
- DMA-148

| Curve     | ka (1/Ms) | kd (1/s) | KD (M)   | Rmax (RU) | Conc (M) |
|-----------|-----------|----------|----------|-----------|----------|
| Cycle: 16 | 1054      | 0.004564 | 4.289E-6 | 7.568     | 3.750E-6 |
| Cycle: 17 | 7.5 µM    |          |          |           |          |
| Cycle: 18 | 15 µM     |          |          |           |          |
| Cycle: 19 | 30 µM     |          |          |           |          |
| Cycle: 20 | 60 µM     |          |          |           |          |
| Cycle: 21 | 15 µM     |          |          |           |          |

| Quality Control | Report | Residuals | Parameters |
|-----------------|--------|-----------|------------|
|                 |        |           |            |

- Kinetic constant ka is approaching the limits that can be measured by the instrument.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions found.
- Check that sensograms have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

- S38
Figure S22. The SPR sensorgrams, fitting parameters and quality control table of DMA-148 (2 replicates)
- DMA-156

**Figure S23.** The SPR sensorgrams, fitting parameters and quality control table of DMA-156 (1 replicate)
- DMA-164

![Graph showing RU vs time for different cycles with kinetic parameters.]

| Curve   | ka (1/Ms) | kd (1/s) | KD (M)   | Rmax (RU) | Conc (M)  |
|---------|-----------|----------|----------|-----------|-----------|
| Cycle: 4 18.75 µM | 75.61     | 0.00660  | 8.015E-5 | 73.42     | 1.875E-5  |
| Cycle: 5 37.5 µM   |           |          |          |           |           |
| Cycle: 6 75 µM     |           |          |          |           |           |
| Cycle: 7 150 µM    |           |          |          |           |           |
| Cycle: 8 300 µM    |           |          |          |           |           |
| Cycle: 9 18.75 µM  |           |          |          |           |           |

*Quality Control* Report: Residuals Parameters

- Kinetic constant ka is outside the limits that can be measured by the instrument.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (R) found.
- Check that sensograms have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

![Graph showing RU vs time for different cycles with kinetic parameters.]

| Curve   | ka (1/Ms) | kd (1/s) | KD (M)   | Rmax (RU) | Conc (M)  |
|---------|-----------|----------|----------|-----------|-----------|
| Cycle: 30 18.75 µM | 131.10     | 0.002328 | 1.776E-5 | 47.74     | 1.875E-5  |
| Cycle: 31 37.5 µM   |           |          |          |           |           |
| Cycle: 32 75 µM     |           |          |          |           |           |
| Cycle: 33 150 µM    |           |          |          |           |           |
| Cycle: 34 300 µM    |           |          |          |           |           |
| Cycle: 35 75 µM     |           |          |          |           |           |
Figure S24. The SPR sensorgrams, fitting parameters and quality control table of DMA-164 (4 replicates)
Figure S25. The SPR sensorgrams, fitting parameters and quality control table of DMA-180 (1 replicate)
• DMA-186

![Graph showing RU response over time for different cycles with corresponding kinetic parameters and concentrations.]

| Curve   | $k_a$ (1/Ms) | $k_d$ (1/s) | $K_D$ (M) | $R_{max}$ (RU) | Conc (M) |
|---------|--------------|-------------|-----------|----------------|----------|
| Cycle: 23 1 μM | 5.823       | 0.01247     | 0.002142  | 4031           | 1.000E-6 |
| Cycle: 24 5 μM | 5.000E-6     |             |           |                |          |
| Cycle: 25 10 μM | 1.000E-5     |             |           |                |          |
| Cycle: 26 25 μM | 2.500E-5     |             |           |                |          |
| Cycle: 27 50 μM | 5.000E-5     |             |           |                |          |
| Cycle: 28 5 μM  | 5.000E-6     |             |           |                |          |

**Quality Control Report**
- Kinetic constant $k_a$ is outside the limits that can be measured by the instrument.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (B) found.
- Check that reaction scheme has sufficient curature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

![Graph showing RU response over time for different cycles with corresponding kinetic parameters and concentrations.]

| Curve   | $k_a$ (1/Ms) | $k_d$ (1/s) | $K_D$ (M) | $R_{max}$ (RU) | Conc (M) |
|---------|--------------|-------------|-----------|----------------|----------|
| Cycle: 23 1.875 μM | 4.658       | 0.006816    | 0.001420  | 5141           | 1.875E-6 |
| Cycle: 24 3.75 μM | 3.750E-6     |             |           |                |          |
| Cycle: 25 7.5 μM | 7.500E-6     |             |           |                |          |
| Cycle: 26 15 μM | 1.500E-5     |             |           |                |          |
| Cycle: 27 30 μM | 3.000E-5     |             |           |                |          |
| Cycle: 28 7.5 μM | 7.500E-6     |             |           |                |          |
Figure S26. The SPR sensorgrams, fitting parameters and quality control table of DMA-186 (3 replicates)
- DMA-187

### Results Table

| Cycle | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|-------|-----------|----------|--------|-----------|----------|
| 23    | 3.125 μM  | 0.01074  | 1.210E-5 | 887.8     | 20.02    |
| 24    | 6.25 μM   |          |        |           | 3.125E-6 |
| 26    | 25 μM     |          |        |           | 6.250E-6 |
| 27    | 50 μM     |          |        |           | 2.500E-5 |
| 28    | 12.5 μM   |          |        |           | 5.000E-5 |

#### Quality Control
- Kinetic constant ka is approaching the limits that can be measured by the instrument.
- High bulk contributions (R) found.
- Check that sensograms have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

### Additional Table

| Cycle | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|-------|-----------|----------|--------|-----------|----------|
| 30    | 1.875 μM  | 0.01144  | 2.769E-5 | 413.1     | 53.99    |
| 31    | 3.75 μM   |          |        |           | 1.875E-5 |
| 32    | 7.5 μM    |          |        |           | 3.750E-5 |
| 33    | 15 μM     |          |        |           | 7.500E-5 |
| 34    | 30 μM     |          |        |           | 1.500E-5 |
| 35    | 7.5 μM    |          |        |           | 3.000E-5 |

S47
Figure S27. The SPR sensorgrams, fitting parameters and quality control table of DMA-187 (3 replicates)
| Curve          | ka (1/Ms) | kd (1/s) | KD (M)     | Rmax (RU) | Conc (M)    |
|---------------|----------|----------|------------|-----------|-------------|
| Cycle: 30     | 2691     | 0.1180   | 4.386E-5   | 33.27     | 3.125E-6    |
| Cycle: 31     | 2691     | 0.1180   | 4.386E-5   | 33.27     | 6.250E-6    |
| Cycle: 32     | 2691     | 0.1180   | 4.386E-5   | 33.27     | 1.250E-5    |
| Cycle: 33     | 2691     | 0.1180   | 4.386E-5   | 33.27     | 2.500E-5    |
| Cycle: 34     | 2691     | 0.1180   | 4.386E-5   | 33.27     | 5.000E-5    |
| Cycle: 35     | 2691     | 0.1180   | 4.386E-5   | 33.27     | 1.250E-5    |

- ** DMA-190 **
Figure S28. The SPR sensorgrams, fitting parameters and quality control table of DMA-190 (2 replicates)
• DMA-191

**Graph and Table:**

| Curve       | ka (1/Ms) | kd (1/s) | KD (M)    | Rmax (RU) | Conc (M) |
|-------------|-----------|----------|-----------|-----------|----------|
| Cycle: 37 1.5625 μM | 1750      | 0.05895  | 3.369E-5  | 111.2     |          |
| Cycle: 38 3.125 μM  | 1.563E-6  |          |           |           |          |
| Cycle: 39 6.25 μM   | 3.125E-6  |          |           |           |          |
| Cycle: 40 12.5 μM   | 6.250E-6  |          |           |           |          |
| Cycle: 41 25 μM     | 1.250E-5  |          |           |           |          |
| Cycle: 42 6.25 μM   | 2.500E-6  |          |           |           |          |

**Quality Control Report Parameters:**
- Kinetic constant ka is approaching the limit that can be measured by the instrument.
- Kinetic constants were difficult to determine.
- Bulk contributions (R) were not evaluated. The R parameter is set to constant.
- Check that sensors are have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S29. The SPR sensorgrams, fitting parameters and quality control table of DMA-191 (2 replicates)
• DMA-193

![Graph showing response over time](image1)

| Curve    | ka (1/\(\text{Ms}\)) | kd (1/\(\text{s}\)) | KD (\(\text{M}\)) | Rmax (RU) | Conc (\(\text{M}\)) |
|----------|-----------------------|---------------------|-------------------|-----------|---------------------|
| Cycle: 30 1.5625 \(\mu\)M | 537.6 | 0.01667 | 3.102E-5 | 663.2 | 1.563E-6 |
| Cycle: 31 3.125 \(\mu\)M | 1.563E-6 | 3.125E-6 |
| Cycle: 32 6.25 \(\mu\)M | 6.250E-6 | |
| Cycle: 33 12.5 \(\mu\)M | 1.250E-5 | |
| Cycle: 35 6.25 \(\mu\)M | 6.250E-6 | |

![Graph showing response over time](image2)

| Curve    | ka (1/\(\text{Ms}\)) | kd (1/\(\text{s}\)) | KD (\(\text{M}\)) | Rmax (RU) | Conc (\(\text{M}\)) |
|----------|-----------------------|---------------------|-------------------|-----------|---------------------|
| Cycle: 16 1.875 \(\mu\)M | 630.9 | 0.01423 | 2.227E-5 | 952.4 | 1.375E-6 |
| Cycle: 17 3.75 \(\mu\)M | 3.750E-6 | |
| Cycle: 18 7.5 \(\mu\)M | 7.500E-6 | |
| Cycle: 19 15 \(\mu\)M | 1.500E-5 | |
| Cycle: 20 30 \(\mu\)M | 3.000E-5 | |
| Cycle: 21 7.5 \(\mu\)M | 7.500E-6 | |
Figure S30. The SPR sensorgrams, fitting parameters and quality control table of DMA-193 (2 replicates)
- DMA-194

![Graph showing time (s) vs. response () with data points and lines for different concentrations.]

| Curve   | ka (1/Ms) | kd (1/s)  | KD (M)      | Rmax (RU) | Conc (M) |
|---------|-----------|-----------|-------------|-----------|----------|
| Cycle: 2 1 μM  | 1373      | 0.01491   | 1.086E-5    | 32.02     | 1.000E-6 |
| Cycle: 3 5 μM  |           |           |             |           | 5.000E-6 |
| Cycle: 4 10 μM |           |           |             |           | 1.000E-5 |

| Quality Control | Report | Residuals | Parameters |
|----------------|--------|-----------|------------|
| ![Icon](icon1.png) | Kinetic constant ka is approaching the limit that can be measured by the instrument. |
| ![Icon](icon2.png) | Kinetic constants appear to be uniquely determined. |
| ![Icon](icon3.png) | High bulk contributions (R) found. |
| ![Icon](icon4.png) | Check that sensorgrams have sufficient curvature. |
| ![Icon](icon5.png) | Examine the residual plot. Pay attention to systematic and non-random deviations. |

![Graph showing time (s) vs. response () with data points and lines for different cycles and concentrations.]
Figure S31. The SPR sensorgrams, fitting parameters and quality control table of DMA-194 (2 replicates)
• TO-PRO-1

![Graph showing RU response over time for different cycles with associated kinetic constants and concentrations.]

| Curve  | $k_a$ (1/Ms) | $k_d$ (1/s) | $K_D$ (M) | $R_{max}$ (RU) | Conc (M)  |
|--------|--------------|-------------|-----------|----------------|-----------|
| Cycle: 11 0.0375 μM | 5.112E+5 | 0.2935  | 5.741E-7 | 76.28 | 2.75E-8 |
| Cycle: 12 0.075 μM | 5.075 | 0.2759 | 5.661E-7 | 75.0E-8 | 3.00E-8 |
| Cycle: 13 0.15 μM | 5.150 | 0.2670 | 5.581E-7 | 74.5E-8 | 3.00E-8 |
| Cycle: 14 0.3 μM | 5.030 | 0.2590 | 5.501E-7 | 74.0E-8 | 3.00E-8 |
| Cycle: 15 0.675 μM | 5.675 | 0.2510 | 5.421E-7 | 73.5E-8 | 3.00E-8 |

| Quality Control | Report | Residuals | Parameters |
|-----------------|--------|-----------|------------|
| ✓ Kinetic constants are within instrument specifications. |
| ✓ Kinetic constants appear to be uniquely determined. |
| ✓ No significant bulk contributions (BD) found. |
| ✓ Check that sensorgrams have sufficient curvature. |
| ✓ Examine the residual plot. Pay attention to systematic and non-random deviations. |
Figure S32. The SPR sensorgrams, fitting parameters and quality control table of TO-PRO-1 (4 replicates)
- Mitoxantrone

![Graph](image1)

| Curve   | $ka$ $(1/\text{Ms})$ | $kd$ $(1/\text{s})$ | $KD$ $(\text{M})$ | $R_{\text{max}}$ $(\text{RU})$ | Conc $(\text{M})$ |
|---------|----------------------|---------------------|-------------------|-------------------------------|------------------|
| Cycle: 3 0.1875 µM | 7.568E+5             | 0.3557              | 4.700E-7          | 388.0                         | 1.875E-7         |
| Cycle: 4 0.375 µM |                     |                     |                   |                               | 3.750E-7         |
| Cycle: 5 0.75 µM  |                     |                     |                   |                               | 7.500E-7         |
| Cycle: 6 1.5 µM   |                     |                     |                   |                               | 1.500E-6         |
| Cycle: 7 3 µM     |                     |                     |                   |                               | 3.000E-6         |
| Cycle: 8 0.1875 µM|                     |                     |                   |                               | 1.875E-7         |

![Graph](image2)

| Curve   | $ka$ $(1/\text{Ms})$ | $kd$ $(1/\text{s})$ | $KD$ $(\text{M})$ | $R_{\text{max}}$ $(\text{RU})$ | Conc $(\text{M})$ |
|---------|----------------------|---------------------|-------------------|-------------------------------|------------------|
| Cycle: 2 0.2 µM | 7.895E+6             | 1.770               | 2.242E-7          | 185.1                         | 2.000E-7         |
| Cycle: 3 0.4 µM  |                     |                     |                   |                               | 4.000E-7         |
| Cycle: 4 0.6 µM  |                     |                     |                   |                               | 6.000E-7         |
| Cycle: 5 0.8 µM  |                     |                     |                   |                               | 8.000E-7         |
| Cycle: 6 1 µM    |                     |                     |                   |                               | 1.000E-6         |
| Cycle: 7 0.6 µM  |                     |                     |                   |                               | 6.000E-7         |
Figure S33. The SPR sensorgrams, fitting parameters and quality control table of mitoxantrone (3 replicates)
• DPF m1

![Graph Image]

| Curve   | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|---------|-----------|----------|--------|-----------|----------|
| Cycle: 3 0.05 µM | 3.539E+5 | 0.04013 | 1.134E-7 | 30.70     | 5.000E-8 |
| Cycle: 4 0.1 µM  |          |          |        |           | 1.000E-7 |
| Cycle: 5 0.5 µM  |          |          |        |           | 5.000E-7 |
| Cycle: 6 1 µM   |          |          |        |           | 1.000E-6 |
| Cycle: 7 0.5 µM |          |          |        |           | 5.000E-7 |

![Graph Image]

| Curve   | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|---------|-----------|----------|--------|-----------|----------|
| Cycle: 9 0.125 µM | 5.407E+5 | 0.06814 | 1.064E-7 | 29.98     | 1.250E-7 |
| Cycle: 10 0.25 µM |          |          |        |           | 2.500E-7 |
| Cycle: 11 0.5 µM  |          |          |        |           | 5.000E-7 |
| Cycle: 12 1 µM   |          |          |        |           | 1.000E-6 |
| Cycle: 13 2 µM   |          |          |        |           | 2.000E-6 |
| Cycle: 14 0.5 µM |          |          |        |           | 5.000E-7 |
Figure S34. The SPR sensorgrams, fitting parameters and quality control table of DPF m1 (3 replicates)
• DPF p1

| Curve     | ka (1/Ms) | kd (1/s) | Kd (M) | Rmax (RU) | Conc (M) |
|-----------|-----------|----------|--------|-----------|----------|
| Cycle: 16 | 1.040E+6  | 0.1065   | 1.024E-7 | 41.86     | 1.250E-7 |
| Cycle: 17 | 0.125 µM  |          |        |           |          |
| Cycle: 18 | 0.25 µM   |          |        |           |          |
| Cycle: 19 | 0.5 µM    |          |        |           |          |
| Cycle: 20 | 1 µM      |          |        |           |          |
| Cycle: 21 | 2 µM      |          |        |           |          |
| Cycle: 22 | 0.5 µM    |          |        |           |          |

**Quality Control**
- Kinetic constants are within instrument specifications.
- Kinetic constants were difficult to determine.
- High bulk contributions (R) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
### Table

| Curve   | $k_a$ (1/Ms) | $k_d$ (1/s) | $K_D$ (M) | $R_{max}$ (RU) | Conc (M) |
|---------|--------------|-------------|-----------|----------------|----------|
| Cycle: 15 0.05 µM | 1.30E+6     | 0.1104      | 8.492E-8  | 151.4          | 5.000E-8 |
| Cycle: 16 0.1 µM  |             |             |           |                | 1.000E-7 |
| Cycle: 17 0.5 µM  |             |             |           |                | 5.000E-7 |
| Cycle: 18 1 µM   |             |             |           |                | 1.000E-6 |
| Cycle: 19 5 µM   |             |             |           |                | 5.000E-6 |
| Cycle: 20 0.5 µM |             |             |           |                | 5.000E-7 |

### Quality Control

- Kinetic constants are within instrument specifications.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (RI) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S35. The SPR sensorgrams, fitting parameters and quality control table of DPF p1 (4 replicates)
- Furamidine

### Curve Data

| Curve   | ka (1/Ms) | kd (1/s) | KD (M)  | Rmax (RU) | Conc (M)   |
|---------|-----------|----------|---------|-----------|------------|
| Cycle: 15 1.25 µM | 2.838E+5  | 0.6950   | 2.449E-6 | 36.08     | 1.250E-6   |
| Cycle: 16 2.5 µM   |           |          |         |           |            |
| Cycle: 17 5 µM     |           |          |         |           |            |
| Cycle: 18 10 µM    |           |          |         |           |            |
| Cycle: 19 1.25 µM  |           |          |         |           |            |

### Quality Control
- Kinetic constant kd is approaching the limits that can be measured by the instrument.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (RI) found.
- Check that sensors have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S36. The SPR sensorgrams, fitting parameters and quality control table of furamidine (2 replicates)
- Ethidium bromide
Figure S37. The SPR sensorgrams, fitting parameters and quality control table of ethidium bromide (5 replicates)
- H-33258

| Curve       | $ka (1/\text{Ms})$ | $kd (1/\text{s})$ | KD (M)    | Rmax (RU) | Conc (M) |
|-------------|--------------------|------------------|-----------|-----------|----------|
| Cycle: 27 0.2 µM | 3.455E+5         | 1.953            | 5.652E-6 | 35.47     | 2.000E-7 |
| Cycle: 28 1 µM     |                   |                  |           |           | 1.000E-6 |
| Cycle: 29 2.5 µM   |                   |                  |           |           | 2.500E-6 |
| Cycle: 30 3 µM     |                   |                  |           |           | 3.000E-6 |
| Cycle: 31 4 µM     |                   |                  |           |           | 4.000E-6 |
| Cycle: 32 5 µM     |                   |                  |           |           | 5.000E-6 |
| Cycle: 33 1 µM     |                   |                  |           |           | 1.000E-6 |

Quality Control Report Residuals Parameters

- Kinetic constant $kd$ is outside the limits that can be measured by the instrument.
- Kinetic constants cannot be uniquely determined.
- High-bulk contributions (R) found.
- Check that sensors have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S38. The SPR sensorgrams, fitting parameters and quality control table of H-33258 (2 replicates)
• DMA-3k

| Curve    | ka (1/Ms) | kd (1/s)  | KD (M)   | Rmax (RU) | Conc (M)   |
|----------|-----------|-----------|----------|-----------|------------|
| Cycle: 30 3.125 μM | 4975      | 0.6106    | 1.227E-4 | 126.9     | 3.125E-6   |
| Cycle: 31 6.25 μM   |           |           |          |           |            |
| Cycle: 32 12.5 μM   |           |           |          |           |            |
| Cycle: 33 15 μM     |           |           |          |           |            |
| Cycle: 34 25 μM     |           |           |          |           |            |

Quality Control Report Residuals Parameters

- Kinetic constant kd is approaching the limits that can be measured by the instrument.
- Kinetic constants were difficult to determine. Try to immobilize less ligand or increase analyte concentration.
- High bulk contributions (RB) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

| Curve    | ka (1/Ms) | kd (1/s)  | KD (M)   | Rmax (RU) | Conc (M)   |
|----------|-----------|-----------|----------|-----------|------------|
| Cycle: 30 3.125 μM | 2.592E+4  | 0.2797    | 1.079E-5 | 11.57     | 3.125E-6   |
| Cycle: 31 6.25 μM   |           |           |          |           |            |
| Cycle: 32 12.5 μM   |           |           |          |           |            |
| Cycle: 34 25 μM     |           |           |          |           |            |
| Curve          | ka (1/Ms) | kd (1/s) | KD (M)    | Rmax (RU) | Conc (M) |
|---------------|-----------|----------|-----------|-----------|----------|
| Cycle: 30 5 µM| 1.898E+4  | 0.2615   | 1.378E-5  | 8.135     | 5.000E-6 |
| Cycle: 31 10 µM|           |          |           |           | 1.000E-5 |
| Cycle: 32 15 µM|           |          |           |           | 1.500E-5 |
| Cycle: 33 20 µM|           |          |           |           | 2.000E-5 |
| Cycle: 34 25 µM|           |          |           |           | 2.500E-5 |
| Cycle: 35 15 µM|           |          |           |           | 1.500E-5 |

**Quality Control**

- **Kinetic constants are within instrument specifications.**
- **Kinetic constants were difficult to determine.**
- **Bulk contributions (RI) were not evaluated. The RI parameter is set to constant.**
- **Check that sensorgram have sufficient curvature.**
- **Examine the residual plot. Pay attention to systematic and non-random deviations.**

![Graph with RU vs Time](image1)

| Curve          | ka (1/Ms) | kd (1/s) | KD (M)    | Rmax (RU) | Conc (M) |
|---------------|-----------|----------|-----------|-----------|----------|
| Cycle: 30 5 µM| 3.125E+4  | 0.2376   | 7.605E-6  | 8.251     | 5.000E-6 |
| Cycle: 31 10 µM|           |          |           |           | 1.000E-5 |
| Cycle: 32 15 µM|           |          |           |           | 1.500E-5 |
| Cycle: 33 20 µM|           |          |           |           | 2.000E-5 |
| Cycle: 34 25 µM|           |          |           |           | 2.500E-5 |
| Cycle: 35 15 µM|           |          |           |           | 1.500E-5 |
Figure S39. The SPR sensorgrams, fitting parameters and quality control table of DMA-3k (5 replicates)
- DMA-31

| Curve   | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|---------|-----------|----------|--------|-----------|----------|
| Cycle: 9 3.125 µM | 139.6     | 0.02186  | 1.552E-4 | 145.9     | 3.125E-6 |
| Cycle: 10 6.25 µM  |           |          |        |           | 6.250E-6 |
| Cycle: 11 12.5 µM  |           |          |        |           | 1.250E-5 |
| Cycle: 12 15 µM    |           |          |        |           | 1.500E-5 |
| Cycle: 13 25 µM    |           |          |        |           | 2.500E-5 |
| Cycle: 14 12.5 µM  |           |          |        |           | 1.250E-5 |

![Graph showing RU over time for DMA-31 with various concentrations and parameters listed in the table above.](image)

![Graph showing RU over time for another set of conditions with parameters listed in the table below.](image)

| Curve   | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|---------|-----------|----------|--------|-----------|----------|
| Cycle: 9 3.125 µM | 177.3     | 0.02337  | 1.318E-4 | 109.9     | 3.125E-6 |
| Cycle: 10 6.25 µM  |           |          |        |           | 6.250E-6 |
| Cycle: 11 12.5 µM  |           |          |        |           | 1.250E-5 |
| Cycle: 12 15 µM    |           |          |        |           | 1.500E-5 |
| Cycle: 13 25 µM    |           |          |        |           | 2.500E-5 |
| Cycle: 14 12.5 µM  |           |          |        |           | 1.250E-5 |
**Quality Control Report**

- Kinetic constant $k_a$ is outside the limits that can be measured by the instrument.
- Kinetic constants cannot be uniquely determined. Try to immobilize less ligand or increase analyte concentration.
- High bulk contributions (R) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

**Graph:**

- RU (Response Units) vs. Time (s)

**Table:**

| Curve   | $ka$ (1/Ms) | $kd$ (1/s) | $KD$ (M)  | $R_{max}$ (RU) | Conc (M)  |
|---------|-------------|------------|-----------|----------------|-----------|
| Cycle: 9 | 3.125 µM    | 128.9      | 0.01214   | 9.419E-5       | 35.45     |
| Cycle: 10| 6.25 µM     | 128.9      | 0.01214   | 9.419E-5       | 3.125E-6  |
| Cycle: 11| 12.5 µM     | 128.9      | 0.01214   | 9.419E-5       | 6.250E-6  |
| Cycle: 12| 15 µM       | 128.9      | 0.01214   | 9.419E-5       | 1.250E-5  |
| Cycle: 13| 25 µM       | 128.9      | 0.01214   | 9.419E-5       | 1.500E-5  |
| Cycle: 14| 12.5 µM     | 128.9      | 0.01214   | 9.419E-5       | 2.500E-5  |

S79
Figure S40. The SPR sensorgrams, fitting parameters and quality control table of DMA-3I (5 replicates)
• DMA-3u

| Curve          | $k_a$ (1/Hz) | $k_d$ (1/s) | $K_D$ (M) | $R_{max}$ (RU) | $Conc$ (M) |
|----------------|--------------|-------------|-----------|----------------|-------------|
| Cycle: 23 3.125 µM | 184.5        | 0.01936     | 1.077E-4  | 58.46          | 3.125E-5    |
| Cycle: 24 6.25 µM  |              |             |           |                | 6.250E-5    |
| Cycle: 25 12.5 µM  |              |             |           |                | 1.250E-5    |
| Cycle: 26 15 µM    |              |             |           |                | 1.500E-5    |
| Cycle: 27 25 µM    |              |             |           |                | 2.500E-5    |
| Cycle: 28 12.5 µM  |              |             |           |                | 1.250E-5    |
### Curve Data

| Curve       | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|-------------|-----------|----------|--------|-----------|----------|
| Cycle: 23 3.125 μM | 197.0     | 0.02510     | 1.274E-4 | 62.50     | 3.125E-6 |
| Cycle: 24 6.25 μM   |           |           |        |           |          |
| Cycle: 25 12.5 μM    |           |           |        |           |          |
| Cycle: 26 15 μM      |           |           |        |           |          |
| Cycle: 27 25 μM      |           |           |        |           |          |
| Cycle: 28 12.5 μM    |           |           |        |           |          |

### Quality Control

- **red** Kinetic constant ka is outside the limits that can be measured by the instrument.
- **red** Kinetic constants cannot be uniquely determined.
- **yellow** Try to immobilize less ligand or increase analyte concentration.
- **yellow** High bulk contributions (R²) found.
- **blue** Check that sensograms have sufficient curvature.
- **blue** Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S41. The SPR sensorgrams, fitting parameters and quality control table of DMA-3u (4 replicates)
• DMA-3v

| Curve       | ka (1/Ms) | kd (1/s) | KD (M)     | Rmax (RU) | Conc (M) |
|-------------|-----------|----------|------------|-----------|----------|
| Cycle: 37 10 μM | 681.6     | 0.02222  | 3.259E-5   | 10.62     | 1.000E-5 |
| Cycle: 38 15 μM |           |          |            |           | 1.500E-5 |
| Cycle: 39 17.95 μM |         |          |            |           | 1.795E-5 |
| Cycle: 40 20 μM |           |          |            |           | 2.000E-5 |
| Cycle: 41 25 μM |           |          |            |           | 2.500E-5 |
| Cycle: 42 20 μM |           |          |            |           | 2.000E-5 |

### Quality Control
- Kinetic constant $k_a$ is approaching the limits that can be measured by the instrument.
- Kinetic constants cannot be uniquely determined. Try to immobilize less ligand or increase analyte concentration.
- High bulk contribution ($R_B$) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
Figure S42. The SPR sensorgrams, fitting parameters and quality control table of DMA-3v (2 replicates)
• DMA-3q

| Curve | $ka$ (1/Ms) | $kd$ (1/s) | $KD$ (M) | $R_{max}$ (RU) | $Conc$ (M) |
|-------|-------------|------------|----------|----------------|------------|
| Cycle: 2 | 1.5625 μM | 24.52 | 0.01915 | 7.810E-4 | 6591 | 1.563E-6 |
| Cycle: 3 | 3.125 μM | 3 | 6.250E-6 | 1.250E-5 | 1.500E-5 |
| Cycle: 4 | 6.25 μM | 1.5625 μM | 0.01915 | 7.810E-4 | 6591 | 1.563E-6 |
| Cycle: 5 | 12.5 μM | 3 | 6.250E-6 | 1.250E-5 | 1.500E-5 |
| Cycle: 6 | 15 μM | 1.5625 μM | 0.01915 | 7.810E-4 | 6591 | 1.563E-6 |

## Notes:
- Kinetic constant $ka$ is outside the limits that can be measured by the instrument.
- Kinetic constants were difficult to determine. Try to immobilize less ligand or increase analyte concentration.
- Bulk contributions (RI) were not evaluated. The RI parameter is set to constant.
- Check that sensograms have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
| Curve   | ka (1/Ms) | kd (1/s) | KD (M)   | Rmax (RU) | Conc (M) |
|---------|-----------|----------|----------|-----------|----------|
| Cycle: 2 1.5625 µM | 27.17     | 0.01828  | 6.727E-4 | 4540      | 1.563E-6 |
| Cycle: 3 3.125 µM   |           |          |          |           |          |
| Cycle: 4 6.25 µM    |           |          |          |           |          |
| Cycle: 5 12.5 µM    |           |          |          |           |          |
| Cycle: 6 15 µM      |           |          |          |           |          |
| Cycle: 7 6.25 µM    |           |          |          |           |          |

**Quality Control**

- **Report**
  - Kinetic constant ka is outside the limits that can be measured by the instrument.
  - Kinetic constants were difficult to determine. Try to immobilize less ligand or increase analyte concentration.
  - High bulk contributions (R) found.
  - Check that sensograms have sufficient curvature.
  - Examine the residual plot. Pay attention to systematic and non-random deviations.

**Graph**

![Graph showing RU vs Time](image)

**Second Table**

| Curve   | ka (1/Ms) | kd (1/s) | KD (M)   | Rmax (RU) | Conc (M) |
|---------|-----------|----------|----------|-----------|----------|
| Cycle: 2 1.5625 µM | 29.78     | 0.02342  | 7.863E-4 | 4584      | 1.563E-6 |
| Cycle: 3 3.125 µM   |           |          |          |           |          |
| Cycle: 4 6.25 µM    |           |          |          |           |          |
| Cycle: 5 12.5 µM    |           |          |          |           |          |
| Cycle: 6 15 µM      |           |          |          |           |          |
| Cycle: 7 6.25 µM    |           |          |          |           |          |
**Quality Control**  |  **Report**  |  **Residuals**  |  **Parameters**
---|---|---|---

- Kinetic constant $k_a$ is outside the limits that can be measured by the instrument.
- Kinetic constants were difficult to determine.
- Try to immobilize less ligand or increase analyte concentration.
- Bulk contributions (RI) were not evaluated. The RI parameter is set to constant.
- Check that sensograms have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

---

**RU**

**Curve** | $k_a$ (1/Ms) | $k_d$ (1/s) | $K_D$ (M) | $R_{max}$ (RU) | Conc (M)
---|---|---|---|---|---
Cycle: 2 1.5625 μM | 16.72 | 0.01213 | 7.257E-4 | 1.509E+4 | 1.563E-6
Cycle: 3 3.125 μM | | | | | 3.125E-6
Cycle: 4 6.25 μM | | | | | 6.250E-6
Cycle: 5 12.5 μM | | | | | 1.250E-5

---

**Quality Control**  |  **Report**  |  **Residuals**  |  **Parameters**
---|---|---|---

- Kinetic constant $k_a$ is outside the limits that can be measured by the instrument.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (RI) found.
- Check that sensograms have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

---

**RU**
Figure S43. The SPR sensorgrams, fitting parameters and quality control table of DMA-3q (5 replicates)
- DMA-3r

### Table 1

| Curve      | $ka$ (1/Ms) | $kd$ (1/s) | $KD$ (M)   | $R_{max}$ (RU) | Conc (M)   |
|------------|-------------|------------|------------|----------------|------------|
| Cycle: 16  | 3.125 µM    | 53.94      | 0.01487    | 2.756E-4       | 328.9      |
| Cycle: 17  | 6.25 µM     | 3.125E-6   | 6.250E-6   | 1.500E-5       | 1.250E-5   |
| Cycle: 19  | 15 µM       | 1.500E-5   | 1.250E-5   |                |            |
| Cycle: 21  | 12.5 µM     |            |            |                |            |

### Observations

- **Warning:** Kinetic constant $ka$ is outside the limits that can be measured by the instrument.
- **Warning:** Kinetic constants appear to be uniquely determined.
- **Warning:** High bulk contributions ($RI$) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
### Quality Control Report

**Residuals**

- Kinetic constant $k_a$ is outside the limits that can be measured by the instrument.
- Kinetic constants were difficult to determine. Try to immobilize less ligand or increase analyte concentration.
- High bulk contributions (RB) found.
- Check that sensors have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

### Graphs

#### Graph 1
- **Response** vs. **Time**

| Curve   | $k_a$ (1/Ms) | $k_d$ (1/s) | $K_D$ (M) | $R_{max}$ (RU) | Conc (M) |
|---------|--------------|-------------|-----------|----------------|-----------|
| Cycle: 16 3.125 μM | 14.86       | 0.009526    | 6.412E-4  | 4393           | 3.125E-6  |
| Cycle: 17 6.25 μM   |             |             |           |                | 6.250E-6  |
| Cycle: 18 12.5 μM   |             |             |           |                | 1.250E-5  |
| Cycle: 19 15 μM     |             |             |           |                | 1.500E-5  |
| Cycle: 21 12.5 μM   |             |             |           |                | 1.250E-5  |

#### Graph 2
- **Response** vs. **Time**
**Figure S44.** The SPR sensorgrams, fitting parameters and quality control table of DMA-3r (4 replicates)
- Thiazole orange

| Curve  | ka (1/Ms) | kd (1/s) | KD (M)   | Rmax (RU) | Conc (M) |
|--------|-----------|----------|----------|-----------|----------|
| Cycle: 2 0.05 μM | 1.542E+5 | 1.208 | 7.832E-6 | 261.7     | 5.000E-8 |
| Cycle: 3 0.2 μM  |          |         |          |           | 2.000E-7 |
| Cycle: 4 0.5 μM  |          |         |          |           | 5.000E-7 |
| Cycle: 5 1 μM   |          |         |          |           | 1.000E-6 |
| Cycle: 6 2.5 μM |          |         |          |           | 2.500E-6 |
| Cycle: 7 5 μM   |          |         |          |           | 5.000E-6 |
| Cycle: 8 1 μM   |          |         |          |           | 1.000E-6 |

- Quality Control: Report Residuals Parameters
  - Kinetic constant k_d is approaching the limits that can be measured by the instrument.
  - Kinetic constants appear to be uniquely determined.
  - High bulk contributions (R) found.
  - Check that sensors have sufficient curvature.
  - Examine the residual plot. Pay attention to systematic and non-random deviations.

![Graph showing Thiazole orange response over time]

![Graph showing Thiazole orange response over time]
Figure S45. The SPR sensorgrams, fitting parameters and quality control table of thiazole orange (3 replicates)
- DPF m3

| Curve | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|-------|-----------|----------|--------|-----------|----------|
| Cycle: 2 | 0.3125 μM | 714.1    | 0.01603 | 2.24E-6   | 75.63    |
| Cycle: 3 | 0.625 μM |          |        |           |          |
| Cycle: 4 | 1.25 μM |          |        |           |          |
| Cycle: 5 | 2.5 μM |          |        |           |          |
| Cycle: 6 | 5 μM |          |        |           |          |
| Cycle: 7 | 1.25 μM |          |        |           |          |
Figure S46. The SPR sensorgrams, fitting parameters and quality control table of DPF m3 (2 replicates)
• DPF m9
Figure S47. The SPR sensorgrams, fitting parameters and quality control table of DPF m9 (2 replicates)
• DPF m10

![Graph showing SPR sensorgrams, fitting parameters and quality control table of DPF m10 (1 replicate)](image)

**Figure S48.** The SPR sensorgrams, fitting parameters and quality control table of DPF m10 (1 replicate)
- **DPF p6**

| Curve  | ka (1/Ms) | kd (1/s) | KD (M)  | Rmax (RU) | Conc (M) |
|--------|-----------|----------|---------|-----------|----------|
| Cycle: 23 0.3125 µM | 1.276E+4  | 0.7802   | 5.114E-5 | 701.3     | 3.125E-7 |
| Cycle: 24 0.625 µM  |           |          |         |           | 6.250E-7 |
| Cycle: 25 1.25 µM   |           |          |         |           | 1.250E-6 |
| Cycle: 26 2.5 µM    |           |          |         |           | 2.500E-6 |
| Cycle: 27 5 µM      |           |          |         |           | 5.000E-6 |
| Cycle: 28 1.25 µM   |           |          |         |           | 1.250E-6 |

### Table Notes:
- **kinetic constant kd is approaching the limits that can be measured by the instrument.**
- Kinetic constants were difficult to determine.
- Bulk contributions (F) were not evaluated. The R parameter is set to constant.
- Check that sensograms have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

---

### Table 2:

| Curve  | ka (1/Ms) | kd (1/s) | KD (M)  | Rmax (RU) | Conc (M) |
|--------|-----------|----------|---------|-----------|----------|
| Cycle: 23 0.3125 µM | 1.442E+4  | 0.7729   | 5.362E-5 | 1103      | 3.125E-7 |
| Cycle: 24 0.625 µM  |           |          |         |           | 6.250E-7 |
| Cycle: 25 1.25 µM   |           |          |         |           | 1.250E-6 |
| Cycle: 26 2.5 µM    |           |          |         |           | 2.500E-6 |
| Cycle: 27 5 µM      |           |          |         |           | 5.000E-6 |
| Cycle: 28 1.25 µM   |           |          |         |           | 1.250E-6 |
Figure S49. The SPR sensorgrams, fitting parameters and quality control table of DPF p6 (2 replicates)
- **DPF p15**

![Graph](image1)

| Curve     | ka (1/Ms) | kd (1/s) | KD (M)  | Rmax (RU) | Conc (M) |
|-----------|-----------|----------|---------|-----------|----------|
| Cycle: 30 | 1.850E+5  | 0.2387   | 1.291E-6| 53.53     | 3.125E-7 |
| Cycle: 31 | 0.3125 µM |          |         |           | 6.250E-7 |
| Cycle: 32 | 1.25 µM   |          |         |           | 1.250E-6 |
| Cycle: 33 | 2.5 µM    |          |         |           | 2.500E-6 |
| Cycle: 34 | 5 µM      |          |         |           | 5.000E-6 |
| Cycle: 35 | 1.25 µM   |          |         |           | 1.250E-6 |

**Quality Control Report Parameters**
- Kinetic constants are within instrument specifications.
- Kinetic constants were difficult to determine.
- Bulk contributions (RI) were not evaluated. The RI parameter is set to constant.
- Check that sensograms have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

![Graph](image2)

| Curve     | ka (1/Ms) | kd (1/s) | KD (M)  | Rmax (RU) | Conc (M) |
|-----------|-----------|----------|---------|-----------|----------|
| Cycle: 9  | 1.832E+5  | 0.1956   | 1.068E-6| 65.58     | 3.125E-7 |
| Cycle: 10 | 0.3125 µM |          |         |           | 6.250E-7 |
| Cycle: 11 | 1.25 µM   |          |         |           | 1.250E-6 |
| Cycle: 12 | 2.5 µM    |          |         |           | 2.500E-6 |
| Cycle: 14 | 1.25 µM   |          |         |           | 1.250E-6 |
**Quality Control**
- Kinetic constants are within instrument specifications.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (R²) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.

| Curve       | ka (1/Ms) | kd (1/s) | KD (M)  | Rmax (RU) | Conc (M) |
|-------------|-----------|----------|---------|-----------|----------|
| Cycle: 37   | 1.971E+5  | 0.2122   | 1.077E-6| 75.76     |          |
| Cycle: 38   | 0.625 M   |          |         |           | 3.125E-7 |
| Cycle: 39   | 1.25 M    |          |         |           | 6.250E-7 |
| Cycle: 40   | 2.5 M     |          |         |           | 1.250E-6 |
| Cycle: 42   | 1.25 M    |          |         |           | 2.500E-6 |

**Quality Control**
- Kinetic constants are within instrument specifications.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (R²) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
**Figure S50.** The SPR sensorgrams, fitting parameters and quality control table of DPF p15 (4 replicates)

| Curve | ka (1/Ms) | kd (1/s) | KD (M) | Rmax (RU) | Conc (M) |
|-------|-----------|----------|--------|-----------|----------|
| Cycle: 30 0.3125 µM | 1.074E+5 | 0.1080 | 1.006E-6 | 97.84 | 3.125E-7 |
| Cycle: 31 0.625 µM |          |         |        |           | 6.250E-7 |
| Cycle: 32 1.25 µM |          |         |        |           | 1.250E-6 |
| Cycle: 33 2.5 µM |          |         |        |           | 2.500E-6 |
| Cycle: 35 1.25 µM |          |         |        |           | 1.250E-6 |

| Quality Control | Report | Result | Parameters |
|-----------------|--------|--------|------------|
| ✔️  | ✔️  | ✔️  | ✔️  |

- Kinetic constants are within instrument specifications.
- Kinetic constants appear to be uniquely determined.
- High bulk contributions (if) found.
- Check that sensorgrams have sufficient curvature.
- Examine the residual plot. Pay attention to systematic and non-random deviations.
- Acridine orange
Figure S51. The SPR sensorgrams, fitting parameters and quality control table of acridine orange (2 replicates)
• **DPF m2**

Figure S52. The SPR sensorgrams, fitting parameters and quality control table of DPF m2 (1 replicate)
Figure S53. The SPR sensorgrams, fitting parameters and quality control table of DPF p2 (1 replicate)
Figure S54. The SPR sensorgrams, fitting parameters and quality control table of DPF p5 (1 replicate)
- DPF p8

(Heterogeneous Ligand binding mode)

- Poor 1:1 fit (high Chi\(^2\))
- Kinetics at limit of detection
- High U-value
- Fast on and off rate constants from Heterogeneous Ligand probably more representative
- Binding parameters were the average of the two binding processes

| Heterogeneous Ligand | k\(_{a1}\) (1/M/s) | k\(_{d1}\) (1/s) | K\(_{D1}\) (M) | k\(_{a2}\) (1/M/s) | k\(_{d2}\) (1/s) | K\(_{D2}\) (M) | rmax1 [RU] | rmax2 [RU] | Conc [M] | tc | Flow [µl/min] | K\(_{D}\) [RU/M] | R [RU] | Chi\(^2\) [RU] | U-value |
|----------------------|------------------|----------------|-------------|------------------|----------------|-------------|-----------|-----------|----------|----|---------------|--------------|--------|----------------|---------|
| Cycle: 26 0.025 µM   | 1.80E+05         | 0.01035        | 5.75E-08    | 4.04E+04        | 0.04786        | 1.98E-06    | 87.71     | 387.3    | 1.61E+09 |    | 5.93E+09      | -1.277       |        |                | 15.4 N/A |
| Cycle: 27 0.05 µM    | 5.50E-08         | 5.93E+09       | 2.832       | 5.50E-08        | 5.93E+09       | 2.832       | 5.93E+09  | 0.8599   |        |    |               |              |        |                |         |
| Cycle: 28 0.1 µM     | 1.00E-07         | 5.93E+09       | 2.784       | 1.00E-07        | 5.93E+09       | 2.784       | 5.93E+09  | 0.432    |        |    |               |              |        |                |         |
| Cycle: 29 0.2 µM     | 2.00E-07         | 5.93E+09       | 0.432       | 2.00E-07        | 5.93E+09       | 0.432       | 5.93E+09  | 0.8599   |        |    |               |              |        |                |         |
| Cycle: 30 0.4 µM     | 4.00E-07         | 5.93E+09       | 0.8599      | 4.00E-07        | 5.93E+09       | 0.8599      | 5.93E+09  | 0.8599   |        |    |               |              |        |                |         |
| Cycle: 31 0.1 µM     | 1.00E-07         | 5.93E+09       | 2.001       | 1.00E-07        | 5.93E+09       | 2.001       | 5.93E+09  | 0.432    |        |    |               |              |        |                |         |

Figure S55. The SPR sensorgrams, fitting parameters and quality control table of DPF p8 (1 replicate)
• DPF p13

(Heterogeneous Ligand binding mode)

• Removed 0.0125 µM curve from analysis due to noise
• Poor 1:1 fit (high \(\text{Chi}^2\))
• Heterogeneous Ligand probably more representative
• Binding parameters were the average of the two binding processes

| Heterogeneous Ligand | \(k_{\text{cat}}\) (1/Ms) | \(k_{\text{off}}\) (1/s) | \(K_D\) (M) | \(k_{\text{cat}}\) (1/Ms) | \(k_{\text{off}}\) (1/s) | \(K_D\) (M) | Rmax1 (RU) | Rmax2 (RU) | Conc (M) | \(IC_{50}\) | Flow (uL/min) | \(t_{\text{Rt}}\) (RU/Ms) | R (RU) | CHF (\%) | U-value |
|----------------------|----------------|----------------|------------|----------------|----------------|------------|------------|------------|---------|----------|----------------|----------------|--------|---------|---------|
| Cycle: 3A 0.025 µM  | 5.623          | 0.004204      | 7.48E-07   | 6.020         | 0.004206      | 6.09E-07   | 559.2      | 557.9      | 6.35E+07 | 50       | 2.34E+08     | 0.6847           | 14.3 | N/A     |  
| Cycle: 3B 0.05 µM   | 5.008          | 0.005204      | 7.48E-07   | 6.020         | 0.004206      | 6.09E-07   | 559.2      | 557.9      | 6.35E+07 | 50       | 2.34E+08     | 0.6847           | 14.3 | N/A     |  
| Cycle: 3C 0.1 µM    | 5.008          | 0.005204      | 7.48E-07   | 6.020         | 0.004206      | 6.09E-07   | 559.2      | 557.9      | 6.35E+07 | 50       | 2.34E+08     | 0.6847           | 14.3 | N/A     |  
| Cycle: 3D 0.05 µM   | 5.008          | 0.005204      | 7.48E-07   | 6.020         | 0.004206      | 6.09E-07   | 559.2      | 557.9      | 6.35E+07 | 50       | 2.34E+08     | 0.6847           | 14.3 | N/A     |  

Figure S56. The SPR sensorgrams, fitting parameters and quality control table of DPF p13 (1 replicate)
• DMZ m3

![Sensorgrams](image)

| Quality Control | Report       | Residuals | Parameters |
|-----------------|--------------|-----------|------------|
| ✓ Kinetic constants are within instrument specifications. | 🚨 Kinetic constants cannot be uniquely determined. Try to immobilize less ligand or increase analyte concentration. | 🔄 High bulk contributions (R) found. | ⚠ Check that sensorgrams have sufficient curvature. |
| 🕊 Examine the residual plot. Pay attention to systematic and non-random deviations. |

| Cycle | Volume | kD (RU/s) | Kd (mM) | Koff (s⁻¹) | Flow (ul/min) | Rmax (RU) | τc | Experiment | RU (RMS) | R² | Adj R² | F-value | P-value |
|-------|--------|-----------|---------|------------|---------------|-----------|----|------------|----------|-----|--------|---------|---------|
| 5     | 0.15 µL| 1.25E-07  | 3.1E+08 | 1.7E+01    | 50            | 1.17E+09  | 3.379| 26         | 0.262    |     |        |         |         |
| 6     | 0.25 µL| 2.50E-07  | 2.0E+08 | 4.0E+01    | 50            | 1.17E+09  | 5.337| 26         | 0.262    |     |        |         |         |
| 8     | 1 µL   | 1.00E+06  | 1.1E+07 | 1.1E+09    | 50            | 8.415     |     | 26         | 0.262    |     |        |         |         |
| 9     | 2 µL   | 2.00E+06  | 1.1E+07 | 1.1E+09    | 50            | 1.39E+01  |     | 26         | 0.262    |     |        |         |         |
| 10    | 0.5 µL | 5.00E+06  | 1.1E+07 | 1.1E+09    | 50            | 6.29E+08  |     | 26         | 0.262    |     |        |         |         |

**Figure S57.** The SPR sensorgrams, fitting parameters and quality control table of DMZ m3 (1 replicate)
Figure S58. The SPR sensorgrams, fitting parameters and quality control table of DMZ p8 (1 replicate)
Figure S59. The SPR sensorgrams, fitting parameters and quality control table of DMZ p13 (1 replicate)
Section D. QSAR modeling

1. Descriptor calculation

\[ \frac{N_1}{N_0} = e^{\frac{-\Delta E}{RT}} \]  \hspace{1cm} \text{Equation S1}

Equation S1 was used to calculate the ratio of two different molecular conformations, where \( \frac{N_1}{N_0} \) is the ratio of the number of molecules in the relative energy states, \( \Delta E \) is the energy difference between \( N_0 \) and \( N_1 \) (3 kcal/mol), \( R \) is the ideal gas constant (0.00198588 kcal/K mol), and \( T \) is the temperature (295 K).

\[ A = \frac{\sum_i A_i e^{\frac{E_i}{k_B T}}}{\sum_i e^{\frac{E_i}{k_B T}}} \]  \hspace{1cm} \text{Equation S2}

For a specific descriptor \( A \), Equation S2 was used as Boltzmann average method to account for multiple conformations of a molecule and give the final descriptor value, where \( A_i \) is the descriptor value of conformation \( i \), \( E_i \) is the energy of conformation \( i \), \( k_B \) is the Boltzmann constant, and \( T \) is the temperature.
2. Methods and scripts

Descriptor refinement (performed on MATLAB (R2020a), use KD data as an example)

- load('KDdata.mat'); % this matrix contains the 1st row as the index of the variables names, 0 for response variable, here is lnKD
- % find features with constant entry>=80%, delete such features resulting new dataset called: data_nonconst
- data=KDdata;
- for i=2:size(data,2)
  - Y(i)=max(sum(data(:,i)==data(:,i)'));
- end
- idx_const=find(Y(:)>=0.8*(size(data,1)-1));
- data_nonconst=data;
- data_nonconst(:,idx_const)=[];

- % find multicolinearity (abs(rho)>0.95) between features, delete ones with more than 1 multicolinearity, based on the max number of multicolinearity, saved the refined data in the data_refine.
- data_refine=data_nonconst;
- cor=corrcoef(data_refine(2:size(data_refine,1),2:size(data_refine,2)));
- cor=abs(cor);
- [a,b]=find(cor>0.95);
- A=[b,a];
- id=find(b>=a);
- A(id,:) = [];
- uni=unique(A(:,1));
- num=zeros(size(uni,1),1);
- for i=1:size(uni,1)
  - idx=find(uni(i)==A(:,1));
  - num(i)=size(idx,1);
- end
- n=0;
- while max(num)>1
  - id_max=find(num==max(num));
  - if size(id_max,1)>1
    - id_max = id_max(1,1);
  - else
    - id_max=id_max;
  - end
  - del_col=find(A(:,1)==uni(id_max));
  - id_del=A(del_col,2);
• data_refine(:,id_del+1)=[];
• cor=corrcoef(data_refine(2:size(data_refine,1),2:size(data_refine,2)));
• cor=abs(cor);
• [a,b]=find(cor>0.95);
• A=[b,a];
• id=find(b>=a);
• A(id,:) = [];
• uni=unique(A(:,1));
• num=zeros(size(uni,1),1);
• for i=1:size(uni,1)
    • idx=find(uni(i)==A(:,1));
    • num(i)=size(idx,1);
• end
• n=n+1; % record how many steps take to complete this task
• end

% in a pair of multicorrelation, delete the one with lower correlation to the y variable
• m=0;
• while size(A,1)>0
    • cor1=abs(corrcoef(data_refine(:,1),data_refine(:,A(1,1)+1)));
    • cor1=cor1(1,end);
    • cor2=abs(corrcoef(data_refine(:,1),data_refine(:,A(1,2)+1)));
    • cor2=cor2(1,end);
    • if cor1>=cor2
        • id_del=A(1,2);
        • data_refine(:,id_del+1)=[];
    • else
        • id_del=A(1,1);
        • data_refine(:,id_del+1)=[];
    • end
    • cor=corrcoef(data_refine(2:size(data_refine,1),2:size(data_refine,2)));
    • cor=abs(cor);
    • [a,b]=find(cor>0.95);
    • A=[b,a];
    • id=find(b>=a);
    • A(id,:) = [];
    • m=m+1; % record how many steps take to complete this task
• end

save('KD_data_refine.mat','data_refine');
Representative data splitting by Kennard-Stone algorithm and PCA (performed on RStudio v1.4.1717)

```r
# load data
data <- read.csv('KD_refine.csv')

# create trainingset and testset id using kenStone on euclidian distance
library(prospectr)
xspace <- data[, -1]
ks <- kenStone(as.matrix(xspace), k = 12, metric = "mahal", pc = 0.99, .center = TRUE, .scale = FALSE)
ks$test
trainid <- ks$test

# assign testset and trainingset
trainingset <- data[trainid,]
testset <- data[-trainid,]

x_train <- as.matrix(trainingset[-1])
y_train <- data.matrix(trainingset[1])
x <- x_train
y <- y_train
x_test <- as.matrix(testset[-1])
y_test <- as.matrix(testset[1])

data_pca <- data

data_pca$lnKD[trainid] = 0
data_pca$lnKD[-trainid] = 1

pc <- prcomp(data_pca[, -1], scale. = TRUE)
summary(pc)
plot(pc, type="lines")

library(rgl)
library(ggplot2)
library(ggfortify)
library(magrittr)

# design figure frame and axis tick
tick_frame <-
data.frame(ticks = seq(-20, 20, length.out = 5),
      zero = 0) %>%
    subset(ticks != 0)

lab_frame <- data.frame(lab = seq(-20, 20),
      zero = 0) %>%
    subset(lab != 0)

tick_sz <- (tail(lab_frame$lab, 1) - lab_frame$lab[1]) / 128

pc_plot <- cbind(data_pca[, 1], pc$x)

# PLOT ----
ggplot(pc_plot, aes(x = pc_plot[, 2], y = pc_plot[, 3])) + labs(x = 'PC1 (29.93%)', y = 'PC2 (20.81%)') +
```

S119
# y axis line
geom_segment(x = 0, xend = 0, y = lab_frame$lab[1], yend = tail(lab_frame$lab, 1), size = 1.5) +

# x axis line
geom_segment(y = 0, yend = 0, x = lab_frame$lab[1], xend = tail(lab_frame$lab, 1), size = 1.5) +

# x ticks
geom_segment(data = tick_frame, aes(x = ticks, xend = ticks, y = zero, yend = zero + tick_sz), size = 1.5) +

# y ticks
geom_segment(data = tick_frame, aes(x = zero, xend = zero + tick_sz, y = ticks, yend = ticks), size = 1.5) +

# labels
geom_text(data = tick_frame, aes(x = ticks, y = zero, label = ticks), vjust = 1.5, size = 6) +
geom_text(data = tick_frame, aes(x = zero, y = ticks, label = ticks), hjust = 1.5, size = 6) +

# legends
scale_color_discrete(name = "dataset", labels = c("Trainingset", "Testset")) +

# THE DATA POINT
geom_point(aes(color = factor(V1)), size = 4, alpha = .6) +
scale_color_manual(labels = c("Training set", "Test set"), values = c("dodgerblue", "red2")) +

# title
ggtitle("Test set molecules in 2D chemical space") +

theme_bw() +
theme(panel.border = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(), axis.ticks.y = element_blank(), axis.text.y = element_blank()) +
theme(axis.title = element_text(size = 22, face = "bold")) +
theme(plot.title = element_text(hjust = 0.5)) +
theme(plot.title = element_text(size = 30, face = "bold")) +
theme(legend.title = element_blank(), legend.text = element_text(color = "black", size = 20, face = "bold")) +
theme(legend.position = "none")

ggsave("datasplit.tiff", units = "in", width = 6, height = 6, dpi = 600)
Descriptor selection by lasso and model selection (performed on RStudio v1.4.1717)

# load data
data <- read.csv("KD_refine.csv")

# Create the evaluation function: eval_results, which contains RMSE and Rsquare
eval_results <- function(true, predicted, df) {
  SSE <- sum((predicted - true)^2)
  SST <- sum((true - mean(true))^2)
  R_square <- 1 - SSE / SST
  RMSE <- sqrt(SSE/nrow(df))

  # Model performance metrics
  data.frame(RMSE = RMSE, Rsquare = R_square)
}

# create trainingset and testset id using kenStone on Mahalanobis distance
library(prospectr)
xspace <- data[, -1]
ks <- kenStone(as.matrix(xspace), k=12, metric = "mahal", pc=0.99, .center = TRUE, .scale = FALSE)
ks$test
trainid <- ks$test

# assign testset and trainingset
trainingset <- data[trainid,]
testset <- data[-trainid,]

x_train <- as.matrix(trainingset[-1])
y_train <- data.matrix(trainingset[1])
x <- x_train
y <- y_train
x_test <- as.matrix(testset[-1])
y_test <- as.matrix(testset[1])

# lasso regression
library(glmnet)
set.seed(1)
lambdas <- 10^seq(2, -6, length = 100)

# use cv.glmnet to find the best lambda for lasso from 5-fold cv
lasso_reg <- cv.glmnet(x_train, y_train, alpha = 1, lambda = lambdas, standardize = TRUE, nfolds = 5)
plot(lasso_reg)

# plot the shrinkage graph with multiple lambda values
lasso_model <- glmnet(x_train, y_train, alpha = 1, nlambda =100,standardize = TRUE)
print(lasso_model)
pl <- plot(lasso_model,xvar="lambda",label = T, lwd=4,cex.lab=2,cex.axis=2,xlim = c(-4.5,0.5), ylim=c(-20,20))
# chose the lambda with lowest mean-squared error from cv
lambda_best_lasso <- lasso_reg$lambda.min

# build the lasso regression model using selected descriptors
lasso_model <- glmnet(x_train, y_train, alpha = 1, lambda = lambda_best_lasso, standardize = TRUE)
summary(lasso_model)

# find the non-zero coefficients and their names
lasso.coef <- predict(lasso_model, type = "coefficients")
lasso.coef[lasso.coef != 0]
lasso_nonzerocoef <- predict(lasso_model, type = "nonzero")
lasso_nonzerocoef

colnames(data[, lasso_nonzerocoef$s0 + 1])

# model evaluation on lasso model using all non-zero descriptors
lasso_fittings <- predict(lasso_model, s = lambda_best_lasso, newx = x_test)
lasso_predictions <- predict(lasso_model, s = lambda_best_lasso, newx = x_test)
eval_results(y_test, lasso_predictions, testset)
eval_results(y_train, lasso_fittings, trainingset)

# exhaustively search for all combinations
# m = number of features in the model, data_step contains all non-zero descriptor candidates, "results" summarizes all results
data_step <- trainingset[, append(lasso_nonzerocoef$s0 + 1, 1, 0)]
m <- 3
idx <- combn(rep(1: (length(data_step) - 1)), m)
results <- NULL
for (i in 1: ncol(idx)) {

data_exhau <- data_step[, append(idx[, i] + 1, 1, 0)]
mdl_exhau <- lm(lnKD ~ ., data = data_exhau)

predict <- predict(mdl_exhau, newdata = testset)
fitted <- mdl_exhau$fitted.values
a <- eval_results(testset$lnKD, predict, testset)
b <- eval_results(trainingset$lnKD, fitted, trainingset)

result <- data.frame(test = a,
  train = b)
results <- rbind(results, result)
}

# idrows find all candidates with top performance, and print out the model summary for statistical significance check
idrows <- which(results$test.Rsquare >= 0.7 & results$train.Rsquare >= .7)

for (val in idrows) {
  data_exhau <- data_step[, append(idx[, val] + 1, 1, 0)]
  mdl_exhau <- lm(lnKD ~ ., data = data_exhau)
s <- summary.mdl_exhau
print(s)
print(val)
cat("R2_test:", results[val,2])
}

# plot the curve for the top model
library(ggplot2)
# load the model
mdl <- lm(formula = "lnKD~1+POE_VSA_POS+vsurf_DW12+vsa_other+vsurf_ID3", 
data = trainingset)
summary(mdl)
predict <- predict(mdl, newdata = data)
id <- numeric(48)
id[-trainid] <- 1
data_plot <- cbind(predict, data$lnKD, id)
colnames(data_plot) <- c("predict", "obs", "id")

ggplot(as.data.frame(data_plot), aes(x=obs, y=predict)) +
  ggtitle(expression("Baseline model of lnK"[D]*"")) +
  xlab(expression("Observed lnK"[D]*"")) + ylab(expression("Predicted lnK"[D]*"")) +
  # THE DATA POINT
  geom_point(aes(color = factor(id)), size = 5, alpha = 1) +
  xlim(min(data$lnKD)-2, max(data$lnKD)+2) +
  ylim(min(data$lnKD)-2, max(data$lnKD)+2) +
  scale_color_manual(labels = c("Training set", "Test set"), values =
c("dodgerblue", "red2")) +
  # title
  theme_bw() +
  theme(axis.ticks.length=unit(.4,"lines")) +
  theme(panel.grid.major = element_blank(), 
        panel.grid.minor = element_blank()) +
  theme(axis.text.y = element_text(size = 20), 
        axis.text.x = element_text(size=20),
        axis.title = element_text(size = 25, face = "bold"), title
        =element_text(size = 25, face = 'bold' ) ) +
  # legend
  theme(legend.title = element_blank()) +
  theme(legend.text = element_text(colour="black", size=20, face="bold")) +
  theme(legend.position = c(0.80, 0.1)) +
  # rec
  theme(panel.background = element_rect(colour = "black", size = 3.5)) +
  # ref line
  geom_abline(intercept = 0, slope = 1, color="black", slope="dashed", size=1.5)

ggsave("KDmdl.tiff", units="in", width=8, height=8, dpi=600)
Ensemble learning-based models (performed on RStudio v1.4.1717)

```r
# load data
data <- read.csv('KD_refine.csv')

# Creat the evaluation function: eval_results, which contained RMSE and Rsquare
eval_results <- function(true, predicted, df) {
  SSE <- sum((predicted - true)^2)
  SST <- sum((true - mean(true))^2)
  R_square <- 1 - SSE / SST
  RMSE <- sqrt(SSE/nrow(df))
  # Model performance metrics
data.frame(    RMSE = RMSE,
    Rsquare = R_square)
}

# create trainingset and testset id using kenStone on Mahalanobis distance
library(prospectr)
xspace <- data[, -1]
ks <- kenStone(as.matrix(xspace), k = 12, metric = "mahal", pc = 0.99, .center = TRUE, .scale = FALSE)
ks$test
trainid <- ks$test

# assign testset and trainingset
trainingset <- data[trainid,]
testset <- data[-trainid,]

x_train <- as.matrix(trainingset[,-1])
y_train <- data.matrix(trainingset[,1])
x <- x_train
y <- y_train
x_test <- as.matrix(testset[,-1])
y_test <- as.matrix(testset[,1])

# build a tree
library(tree)
tree.KD <- tree(lnKD ~ ., data, subset = trainid)
plot(tree.KD)
text(tree.KD)

# evaluate the prediction and fitting
pred_KD <- predict(tree.KD, newdata = testset)
fitted <- predict(tree.KD, , newdata = trainingset)
eval_results(testset$lnKD, pred_KD, testset)
eval_results(trainingset$lnKD, fitted, trainingset)

# use CV to select best size
set.seed(1)
tree.KD_cv <- cv.tree(tree.KD)
plot(tree.KD_cv$size, tree.KD_cv$dev, type = 'b')
prune_KD <- prune.tree(tree.KD, best = 6)
pred_KD <- predict(prune_KD, newdata = testset)
plot(prune_KD)
```

S124
text(prune_KD)
fitted <- predict(prune_KD,,newdata = trainingset)
eval_results(testset$lnKD,pred_KD,testset)
eval_results(trainingset$lnKD,fitted,trainingset)

# bagging: set mtry = 193 in randomForest method
library(randomForest)
set.seed(1)
rf_KD <- randomForest(lnKD~.,data = trainingset,importance = TRUE,ntree = 200,sampsize=24, mtry = 193)
ssummary(rf_KD)
print(rf_KD)

# plot
plot(rf_KD,main ="Averaged OOB error", cex.lab=2,cex.axis=2,cex.main=2, lwd=4, col = "red")
pred_KD <-predict(rf_KD,newdata = testset)
fitted <- predict(rf_KD,,newdata = trainingset)
eval_results(testset$lnKD,pred_KD,testset)
eval_results(trainingset$lnKD,fitted,trainingset)
varImpPlot(rf_KD,main = "Variable importance plot")

# random forest
library(randomForest)
set.seed(1)
rf_KD <- randomForest(lnKD~.,data = trainingset,importance = TRUE,sampsize = 34,ntree = 100,mty=40)
ssummary(rf_KD)
print(rf_KD)
plot(rf_KD,main ="Averaged OOB error", cex.lab=2,cex.axis=2,cex.main=2, lwd=4, col = "red")
pred_KD <-predict(rf_KD,newdata = testset)
fitted <- predict(rf_KD,,newdata = trainingset)
eval_results(testset$lnKD,pred_KD,testset)
eval_results(trainingset$lnKD,fitted,trainingset)
varImpPlot(rf_KD,main = "Variable importance plot")

# plot
predict <- predict(rf_KD,newdata = data)
id <- numeric(48)
id[-trainid] <- 1
data_plot <- cbind(predict,data$lnKD,id)
colnames(data_plot) <- c("predict", "obs","id")
ggplot(as.data.frame(data_plot), aes(x=obs,y=predict)) +
ggtitle(expression("Baseline model of lnK"[D]*"")) +
  xlab(expression("Observed lnK"[D]*"")) + ylab(expression("Predicted
lnK"[D]*"")) +
  # THE DATA POINT
  geom_point(aes(color = factor(id)),size = 5,alpha =1) +
xlim(min(data$lnKD)-2,max(data$lnKD)+2)+
  ylim(min(data$lnKD)-2,max(data$lnKD)+2)+
scale_color_manual(labels = c("Training set", "Test set"), values =
c("dodgerblue", "red2"))+
  # title
theme_bw()
theme(axis.ticks.length=unit(.4,"lines"))
theme(panel.grid.major = element_blank(),
  panel.grid.minor = element_blank())
theme(axis.text.y = element_text(size = 20),
  axis.text.x = element_text(size = 20),
  axis.title = element_text(size = 25, face = 'bold'), title
  =element_text(size = 25, face = 'bold') )+
  # legend
theme(legend.title = element_blank())
theme(legend.text = element_text(colour="black", size=20, face="bold"))
theme(legend.position = c(0.80, 0.1)) +
  # rec
theme(panel.background = element_rect(colour = "black", size = 3.5)) +
  # ref line
ggtitle(expression("Baseline model of lnK"))

library(gbm)
set.seed(1)
boost_KD <- gbm(lnKD~., data=trainingset, distribution = 'gaussian', n.trees=2000, interaction.depth=1,
  shrinkage = 0.01, cv.folds = 5, verbose = TRUE, n.minobsinnode=4, bag.fraction = 0.5 )
summary(boost_KD)
print(boost_KD)
sqrt(min(boost_KD$cv.error))

gbm.perf(boost_KD, method = "cv")
legend(1200, .5, c("OOB(Out Of Bag estimator method)", "CV(Cross Validation method)"), cex=0.8, col=c("black", "green"), lty=1)
pred_KD <- predict(boost_KD, newdata = testset, n.trees = 990)
fitted <- predict(boost_KD, newdata = trainingset, n.trees = 990)
eval_results(testset[,lnKD, pred_KD, testset]
eval_results(trainingset[,lnKD, fitted, trainingset)

# plot
predict <- predict(boost_KD, newdata = data, n.trees = 500)
id <- numeric(48)
id[-trainid] <- 1
data_plot <- cbind(predict, data$lnKD, id)
colnames(data_plot) <- c("predict", "obs", "id")
ggplot(as.data.frame(data_plot), aes(x=obs, y=predict)) +
ggtitle(expression("Baseline model of lnK")) +
  xlab(expression("Observed lnK")) + ylab(expression("Predicted lnK")) +
  # THE DATA POINT
  geom_point(aes(color = factor(id)), size = 5, alpha = .1) +
  xlim(min(data$lnKD)-2, max(data$lnKD)+2) +
  ylim(min(data$lnKD)-2, max(data$lnKD)+2) +
  scale_color_manual(labels = c("Training set", "Test set"), values = c("dodgerblue", "red2")) +
  # title
theme_bw() +
theme(axis.ticks.length = unit(.4, "lines")) +
theme(panel.grid.major = element_blank(),
      panel.grid.minor = element_blank()) +
theme(axis.text.y = element_text(size = 20),
      axis.text.x = element_text(size = 20),
      axis.title = element_text(size = 25, face = 'bold'),
      title = element_text(size = 25, face = 'bold') ) +
  # legend
theme(legend.title = element_blank()) +
theme(legend.text = element_text(colour = "black", size = 20, face = "bold") ) +
theme(legend.position = c(0.80, 0.1)) +
  # rec
theme(panel.background = element_rect(colour = "black", size = 3.5)) +
  # ref line
geom_abline(intercept = 0, slope = 1, colour = "black",
             linetype = "dashed", size = 1.5)
ggsave("gbmmdl.tiff", units = "in", width = 8, height = 8, dpi = 600)
Model assessment: Q-Q plot and Williams plot (performed on RStudio v1.4.1717)

```r
# load data
data <- read.csv('KD_refine.csv')

# create trainingset and testset id using kenStone on Mahalanobis distance
library(prospectr)
xspace <- data[, -1]
ks <- kenStone(as.matrix(xspace), k = 12, metric = "mahal", pc = 0.99, .center = TRUE, .scale = FALSE)
ks$test
trainid <- ks$test

# assign testset and trainingset
trainingset <- data[trainid,]
testset <- data[-trainid,]
x_train <- as.matrix(trainingset[-1])
y_train <- data.matrix(trainingset[1])
x <- x_train
y <- y_train
x_test <- as.matrix(testset[-1])
y_test <- as.matrix(testset[1])

# model gonna be assessed
mdl <- lm(formula = "lnKD~1+PEOE_VSA_POS+vsurf_DW12+vsa_other+vsurf_ID3",
data = trainingset)
summary(mdl)

# plot q-q plot
qqnorm(mdl$residuals, pch = 19, cex = 2.5, col = "blue")
qqline(mdl$residuals, col = "black", lwd = 3, lty = 2)

# Williams plot for lnKD model
library(matlib)
library(ggplot2)
wp_x <- cbInd(data$PEOE_VSA_POS, data$vsurf_DW12, data$vsa_other, data$vsurf_ID3)

h <- diag(wp_x %*% inv(t(wp_x) %*% wp_x) %*% t(wp_x))
stdres_train <- (mdl$residuals-mean(mdl$residuals))/sd(mdl$residuals)
res_test <- predict(mdl, newdata=testset)-testset$lnKD
stdres_test <- (res_test-mean(mdl$residuals))/sd(mdl$residuals)

wp_mt <- matrix(0, 48, 3)
wp_mt[testid, 1] <- 1
wp_mt[, 2] <- h
wp_mt[trainid, 3] <- stdres_train
wp_mt[testid, 3] <- stdres_test

colnames(wp_mt) = c("id", "hatvalue", "stdres")

ggplot(as.data.frame(wp_mt), aes(x=hatvalue, y=stdres)) +
ggtitle(expression("Williams plot: lnK[D]")) +
```
```r
xlab(expression("Leverage")) + ylab(expression("Standardized residuals")) +
# THE DATA POINT
geom_point(aes(color = factor(id)), size = 5, alpha = 1) +
xlim(0, 0.8) +
ylim(-4, 4) +
scale_color_manual(labels = c("Training set", "Test set"), values =
c("dodgerblue", "red2")) +

# title
theme_bw() +
theme(axis.ticks.length = unit(.4, "lines")) +
theme(panel.grid.major = element_blank(),
      panel.grid.minor = element_blank()) +
theme(axis.text.y = element_text(size = 20),
      axis.text.x = element_text(size = 20),
      axis.title = element_text(size = 25, face = 'bold'),
      title = element_text(size = 25, face = 'bold')) +

# legend
theme(legend.title = element_blank()) +
theme(legend.text = element_text(colour = "black", size = 20, face = "bold")) +
theme(legend.position = c(0.80, 0.1)) +

# rec
theme(panel.background = element_rect(colour = "black", size = 3.5)) +

# ref line
geom_abline(intercept = 3, slope = 0, color = "black",
            linetype = "dashed", size = 1.5) +
geom_abline(intercept = -3, slope = 0, color = "black",
            linetype = "dashed", size = 1.5) +
geom_vline(xintercept = 3 * 5/36, color = "black",
            linetype = "dashed", size = 1.5)

plot(diag(h), stdred, col = c("blue4"), pch = 19, cex = 2, cex.lab = 2, cex.axis = 2,
     xlim = c(0, 0.5), ylim = c(-4, 4))
```
Predictor stability test (performed on RStudio v1.4.1717)

# load data
data <- read.csv('KD_refine.csv')

# Creat the evaluation function: eval_results, which contains RMSE and Rsquare
eval_results <- function(true, predicted, df) {
  SSE <- sum((predicted - true)^2)
  SST <- sum((true - mean(true))^2)
  R_square <- 1 - SSE / SST
  RMSE = sqrt(SSE/nrow(df))
  # Model performance metrics
data.frame(RMSE = RMSE,
    Rsquare = R_square)
}

# randomize the data splitting 100 times (36:12)
results <- NULL
for (i in 1:100){
  set.seed(i)
  testid <- sample(seq_len(nrow(data)),size=12)

  # assign testset and trainingset
  trainingset <- data[-testid,]
  testset <- data[testid,]
  x_train <- as.matrix(trainingset[-1])
  y_train <- data.matrix(trainingset[1])
  x <- x_train
  y <- y_train
  x_test <- as.matrix(testset[-1])
  y_test <- as.matrix(testset[1])
  mdl <- lm(formula = "lnKD~1+PEOE_VSA_POS+vsa_other+vsurf_DW12+vsurf_ID3",
    data = trainingset)  # using the same descriptors to build the model
  predict <- predict(mdl,newdata = testset)
  fitted <- mdl$fitted.values
  a <- eval_results(testset$lnKD, predict, testset)
  b <- eval_results(trainingset$lnKD, fitted, trainingset)
  result <- data.frame(test=a,
    train=b)
  results<- rbind(results,result)
}

# plot
barplot(results$test.Rsquare,xlim = c(0,i*1.2),ylim=c(-.5,1.5),lwd=3)
abline(h=mean(results$test.Rsquare), col ="Red",lwd = 5,xlim=c(0,i))
text(x = c(0.1*i,0.3*i,0.4*i,0.5*i),
  y = c(1.2,1.2,1.2,1.2),cex = 1.5,
  labels = c("R2_test = ", round(mean(results$test.Rsquare),2), ", +/-",
round(sd(results$test.Rsquare),2)))
barplot(results$train.Rsquare, xlim = c(0,i*1.2), ylim=c(0,1), lwd=3)
abline(h = mean(results$train.Rsquare), col = "Red", lwd = 5, xlim=c(0,1))
text(x = c(0.1*i, 0.3*i, 0.4*i, 0.5*i),
     y = c(0.9, 0.9, 0.9, 0.9), cex = 1.5,
     labels = c("R2_train = ", round(mean(results$train.Rsquare), 2), "+/-",
      round(sd(results$train.Rsquare), 2)))
References:

1. Patwardhan, N. N.; Ganser, L. R.; Kapral, G. J.; Eubanks, C. S.; Lee, J.; Sathyamoorthy, B.; Al-Hashimi, H. M.; Hargrove, A. E., Amiloride as a new RNA-binding scaffold with activity against HIV-1 TAR. *MedChemComm* **2017**, *8* (5), 1022-1036.
2. Patwardhan, N. N.; Cai, Z.; Umuhire Juru, A.; Hargrove, A. E., Driving factors in amiloride recognition of HIV RNA targets. *Org Biomol Chem* **2019**, *17* (42), 9313-9320.
3. Umuhire Juru, A.; Cai, Z.; Jan, A.; Hargrove, A. E., Template-guided selection of RNA ligands using imine-based dynamic combinatorial chemistry. *Chemical Communications* **2020**, *56* (24), 3555-3558.
4. Donlic, A.; Morgan, B. S.; Xu, J. L.; Liu, A.; Roble Jr, C.; Hargrove, A. E., Discovery of Small Molecule Ligands for MALAT1 by Tuning an RNA-Binding Scaffold. *Angewandte Chemie* **2018**, *130* (40), 13426-13431.
5. Donlic, A.; Zafferani, M.; Padroni, G.; Puri, M.; Hargrove, Amanda E., Regulation of MALAT1 triple helix stability and in vitro degradation by diphenylfurans. *Nucleic Acids Research* **2020**, *48* (14), 7653-7664.