

Supporting Information

Chemical Space Exploration of DprE1 Inhibitors Using Cheminformatics and Artificial Intelligence

Sonali Chhabra^{†‡}, Sunil Kumar[†], and Raman Parkesh^{†‡*}

[†] CSIR-Institute of Microbial Technology, Chandigarh, 160036, India

[‡]Academy of Scientific and Innovative Research, Ghaziabad, 201002, India

*E-mail: rparkesh@imtech.res.in. Tel.: +91 172 6665488. Fax: +91 172 2690585.

Table S1. Loading value for the first three principal components of the physicochemical properties of the MIC value dataset.

Table S2. Loading value for the first three principal components of the physicochemical properties of the IC₅₀ value dataset.

Table S3. Comparative principal component analysis (PCA) of physicochemical properties for anti-tuberculosis drugs, DprE1 inhibitors and FDA approved drugs.

Table S4. Comparison of structural similarity, biological activity, and SALI value among the pairs of small molecules of MIC value dataset.

Table S5. Comparison of structural similarity, biological activity, and SALI value among the pairs of small molecules of IC₅₀ value dataset.

Figure S1. Three-dimensional visualization of PCA of physicochemical properties of DprE1 small molecule inhibitors. (A) MIC value dataset for properties: drug-likeness, number of hydrogen bond acceptors (H-acceptors), number of hydrogen bond donors (H-donors), polar

surface area (PSA), relative polar surface area, rotatable bonds, molecular weight, total surface area, lipophilicity (cLogP), and aqueous solubility (cLogS). (B) IC₅₀ value dataset properties: drug-likeness, hydrogen bond acceptor, hydrogen bond donor, polar surface area, relative polar surface area, rotatable bonds, molecular weight, total surface area, lipophilicity (cLogP), and aqueous solubility (cLogS), ligand efficiency (LE), ligand lipophilicity efficiency (LLE) and lipophilicity-corrected ligand efficiency (LELP).

Figure S2. (A) Activity cliff set based on neighborhood similarity relationship of MIC value dataset. Colors indicate p-value, where red color represents high value and blue color represents low value. The size indicates the SALI p-value of SkeleSpheres. (B) SALI plot of compound pairs. X and Y-axis represent activity values being plotted; color indicates the delta activity; higher and lower values indicated by red and blue colors, respectively. The size of the scatter indicates the SALI value. (C) Example of the compound pair (ID: 841 and ID: 860) that represents the activity cliff. The major structural differences between the compounds are highlighted with red circles.

Figure S3. (A) Activity cliff set based on neighborhood similarity relationship of IC₅₀ value dataset. Colors indicate p-value, where red color represents high value and blue color represents low values. The size indicates the SALI p-value of SkeleSpheres. (B) SALI plot of compound pairs. X and Y-axis represent activity values being plotted; color indicates the delta activity; higher and lower values indicated by red and blue colors, respectively. The size of the scatter indicates the SALI value. (C) Example of the compound pair (ID: 134 and ID: 221) that represents the activity cliff. The major structural differences between the compounds are highlighted with red circles.

Figure S4. Core fragments vs activity (p-value) represented by scatter plot. (A) MIC value dataset. (B) IC₅₀ values dataset. The X-axis represents the core fragments while the Y-axis represents the p-value. The color indicates the p-value, with red and blue color indicating higher and lower p-value, respectively.

Figure S5. (A) Representative examples of small molecule inhibitors of DprE1 IC₅₀ dataset depicting the characteristic structural features contributing to SAR. (B) Cumulative response plot of the inhibitor model, showing the relation between the percentage of hits (y-axis) and

percentile (x-axis). (C) Lift curve of the inhibitor model depicting observations from the percentile about the outperformance of the model over a random model. (D) ROC plot of the inhibitor representing percent of hits (y-axis) and false alarms (x-axis). (E) Predicted-actual scatter plot of the inhibitor model.

Figure S6. (A) Representative examples of small molecule inhibitors of DprE1 MIC dataset depicting the characteristic structural features contributing to SAR. (B) Cumulative response plot of the inhibitor model, showing the relation between the percentage of hits (y-axis) and percentile (x-axis). (C) Lift curve of the inhibitor model depicting observations from the percentile about the outperformance of the model over a random model. (D) ROC plot of the inhibitor representing percent of hits (y-axis) and false alarms (x-axis). (E) Predicted-actual scatter plot of the inhibitor model.

Table S1. Loading value for the first three principal components of the physicochemical properties of the MIC value dataset.

Variable Name	PC1	PC2	PC3
Drug likeness	-0.217	0.217	0.559
Hydrogen Bond Acceptors	-0.204	-0.484	0.177
Hydrogen Bond Donors	-0.189	0.0974	0.505
Polar Surface Area	-0.283	-0.478	-0.0566
Relative Polar Surface Area	-0.419	-0.313	-0.161
Rotatable Bonds	0.359	-0.138	0.0164
Total Molecular Weight	0.297	-0.379	0.352
Total Surface Area	0.392	-0.287	0.402
Lipophilicity	0.47	0.0752	-0.114
Aqueous Solubility	-0.167	0.362	0.27

Table S2. Loading value for the first three principal components of the physicochemical properties of the IC₅₀ value dataset.

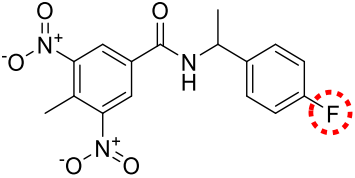
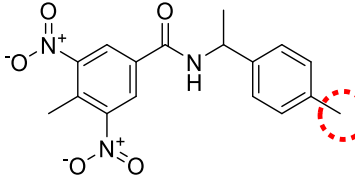
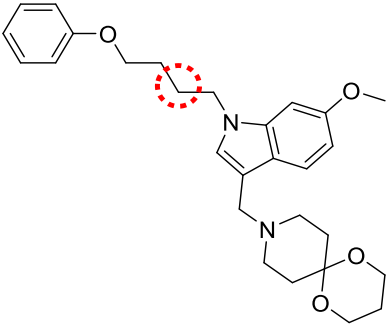
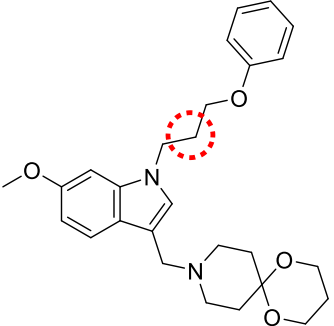
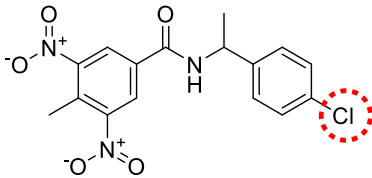
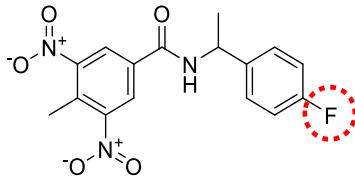
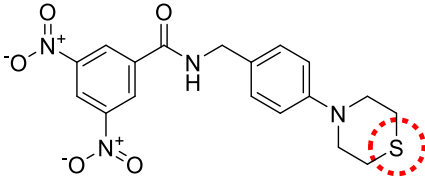
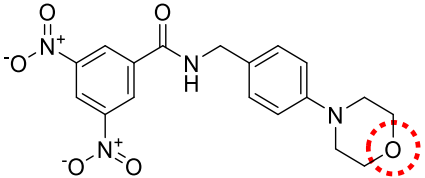
Variable Name	PC1	PC2	PC3
Drug likeness	-0.207	0.15	0.476
Hydrogen Bond Acceptors	-0.252	-0.457	0.0207
Hydrogen Bond Donors	-0.134	-0.021	0.591
LE from IC50 (μ M)	-0.247	0.365	0.0944
LELP from IC50 (μ M)	0.377	-0.0177	0.0201
LLE from IC50 (μ M)	-0.343	0.0342	0.233
Polar Surface Area	-0.285	-0.414	-0.167
Relative Polar Surface Area	-0.327	-0.268	-0.217
Rotatable Bonds	0.263	0.0405	0.0532
Total Molecular Weight	0.186	-0.483	0.338
Total Surface Area	0.276	-0.264	0.381
Lipophilicity	0.367	0.0871	0.0308
Aqueous Solubility	-0.229	0.278	0.145

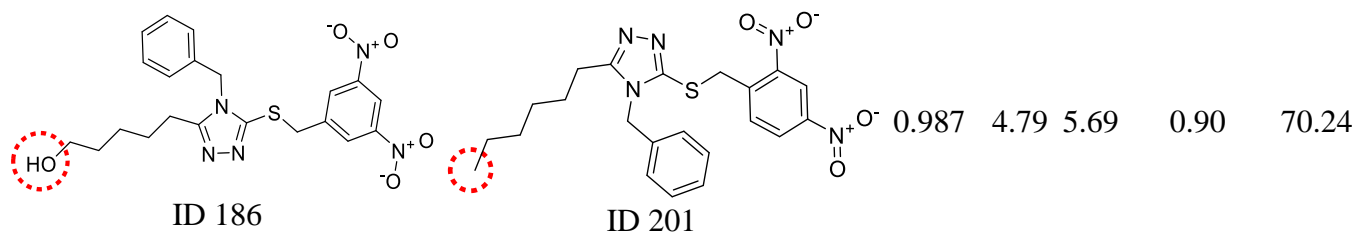
LE: Ligand Efficiency; LELP: Lipophilicity-Corrected Ligand Efficiency; LLE: Ligand Lipophilicity Efficiency

Table S3. Comparative principal component analysis (PCA) of physiochemical properties for anti-tuberculosis drugs, DprE1 inhibitors and FDA approved drugs.

Variable Name	PC1	PC2	PC3
Molecular Weight	0.404	-0.226	-0.089
Lipophilicity (cLogP)	-0.175	-0.521	0.061
Solubility (cLogS)	-0.105	0.509	-0.007
Hydrogen Bond Acceptors	0.433	0.072	0.027
Hydrogen Bonds Donors	0.363	0.188	0.099
Total Surface Area	0.395	-0.265	0.088
Relative Polar Surface Area	0.136	0.490	-0.141
Polar Surface Area	0.424	0.147	-0.010
Drug likeness	-0.027	0.099	0.960
Rotatable Bonds	0.346	-0.195	-0.169
Cumulative Proportion	0.496	0.761	0.864

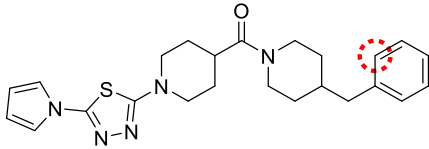
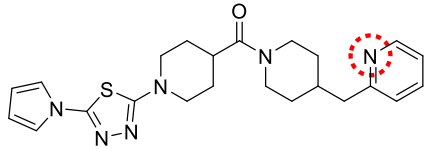
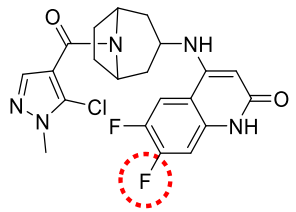
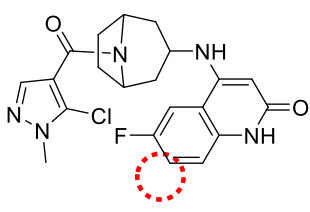
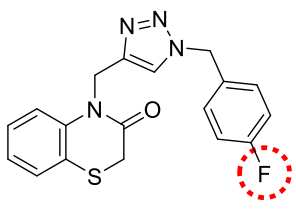
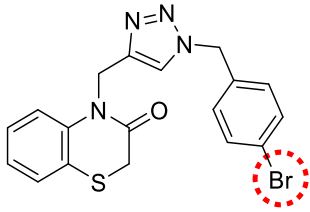
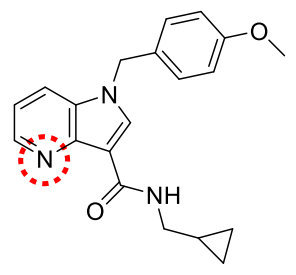
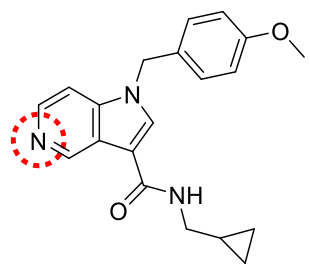
Table S4. Comparison of structural similarity, biological activity, and SALI value among the pairs of small molecules of MIC value dataset.

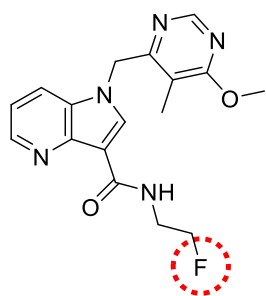
Structure [†] 1	Structure [†] 2	Similarity	Activity* 1	Activity* 2	Delta Activity	SALI#
 <p>ID 841</p>	 <p>ID 860</p>	0.973	7.39	4.93	2.46	94.66
 <p>ID 159</p>	 <p>ID 227</p>	0.988	4.95	5.88	0.92	82.61
 <p>ID 750</p>	 <p>ID 841</p>	0.973	5.25	7.39	2.13	82.15
 <p>ID 513</p>	 <p>ID 603</p>	0.972	7.20	5.09	2.10	77.69



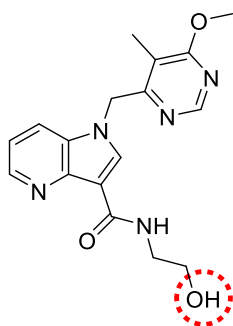
*Activity 1 and Activity 2 are in terms of p-value, #Structure-Activity Landscape Index,
 †Highlighted red circles indicate structural differences.

Table S5. Comparison of structural similarity, biological activity, and SALI value among the pairs of small molecules of IC₅₀ value dataset.

Structure [†] 1	Structure [†] 2	Similarity	Activity* 1	Activity* 2	Delta Activity	SALI#
 <p>ID 134</p>	 <p>ID 221</p>	0.969	7.26	5.58	1.68	55.54
 <p>ID 82</p>	 <p>ID 319</p>	0.931	8.04	4.36	3.67	53.55
 <p>ID 196</p>	 <p>ID 299</p>	0.973	5.90	4.52	1.38	53.14
 <p>ID 71</p>	 <p>ID 187</p>	0.959	8.15	6	2.15	53.07



ID 84



ID 143

0.981 8 7.04 0.95 52.65

*Activity 1 and Activity 2 are in terms of p-value, #Structure-Activity Landscape Index,
†Highlighted red circles indicate structural differences.

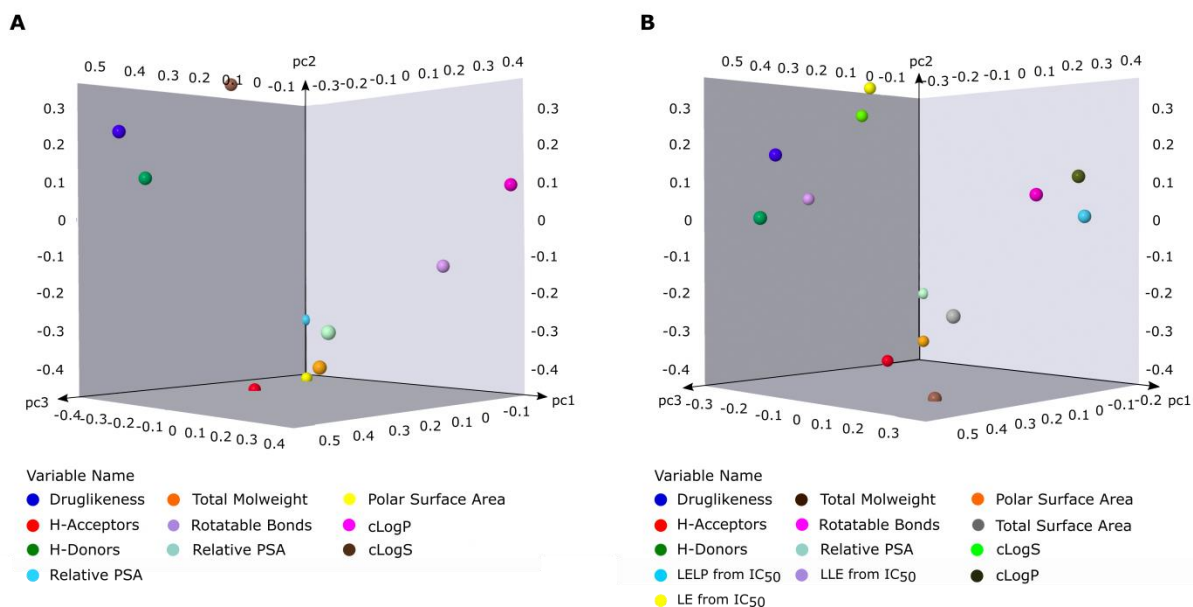


Figure S1. 3D visualization of PCA of physicochemical properties of DprE1 small molecule inhibitors. (A) MIC value dataset for properties: drug-likeness, number of hydrogen bond acceptors (H-acceptors), number of hydrogen bond donors (H-donors), polar surface area (PSA), relative polar surface area, rotatable bonds, molecular weight, total surface area, lipophilicity (cLogP), and aqueous solubility (cLogS). (B) IC₅₀ value dataset properties: drug-likeness, hydrogen bond acceptor, hydrogen bond donor, polar surface area, relative polar surface area, rotatable bonds, molecular weight, total surface area, lipophilicity (cLogP), and aqueous solubility (cLogS), ligand efficiency (LE), ligand lipophilicity efficiency (LLE) and lipophilicity-corrected ligand efficiency (LELP).

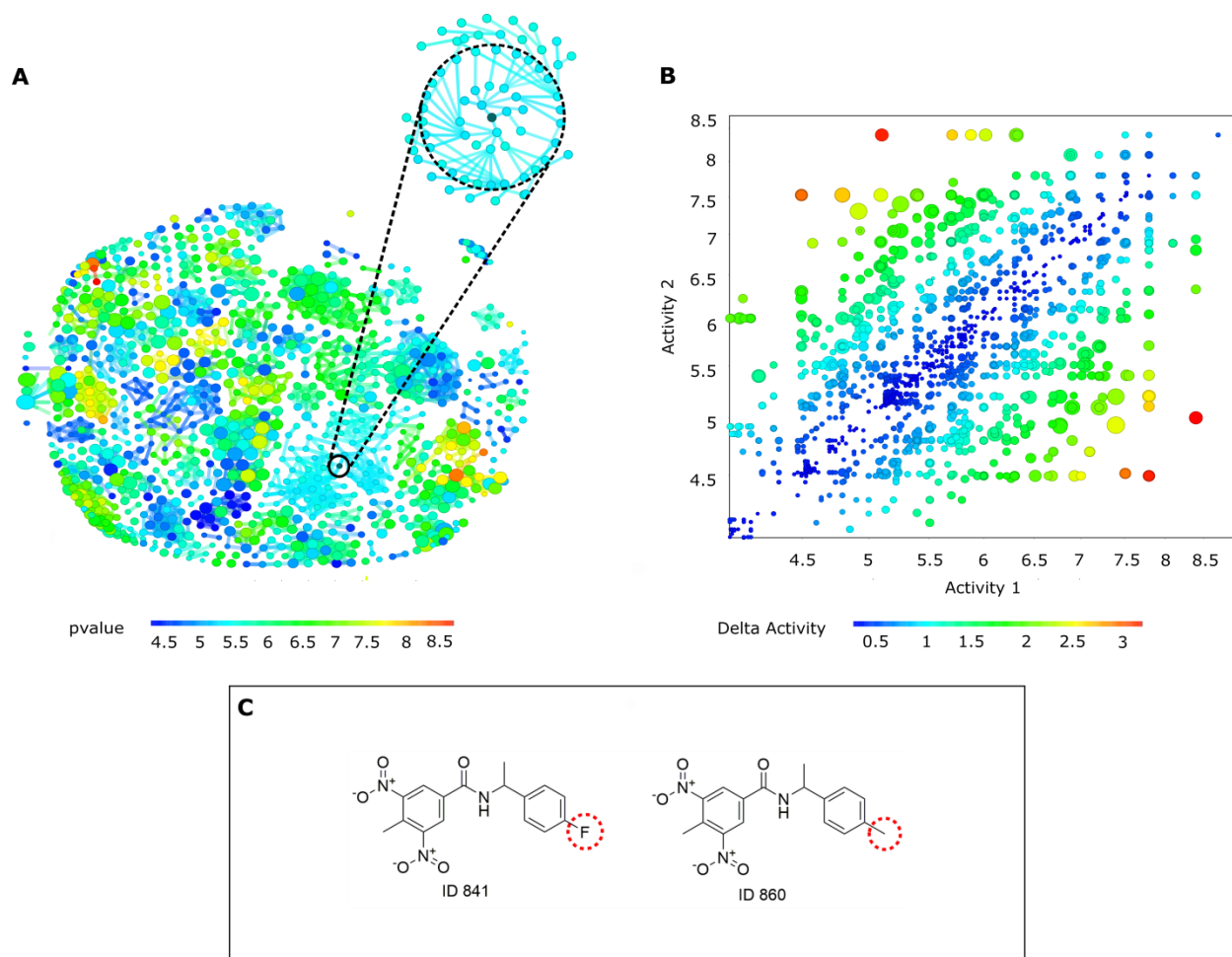


Figure S2. (A) Activity cliff set based on neighborhood similarity relationship of MIC value dataset. Colors indicate p-value, where red color represents high value and blue color represents low value. The size indicates the SALI p-value of SkeleSpheres. (B) SALI plot of compound pairs. X and Y-axis represent activity values being plotted; color indicates the delta activity; higher and lower values indicated by red and blue colors, respectively. The size of the scatter indicates the SALI value. (C) Example of the compound pair (ID: 841 and ID: 860) that represents the activity cliff. The major structural differences between the compounds are highlighted with red circles.

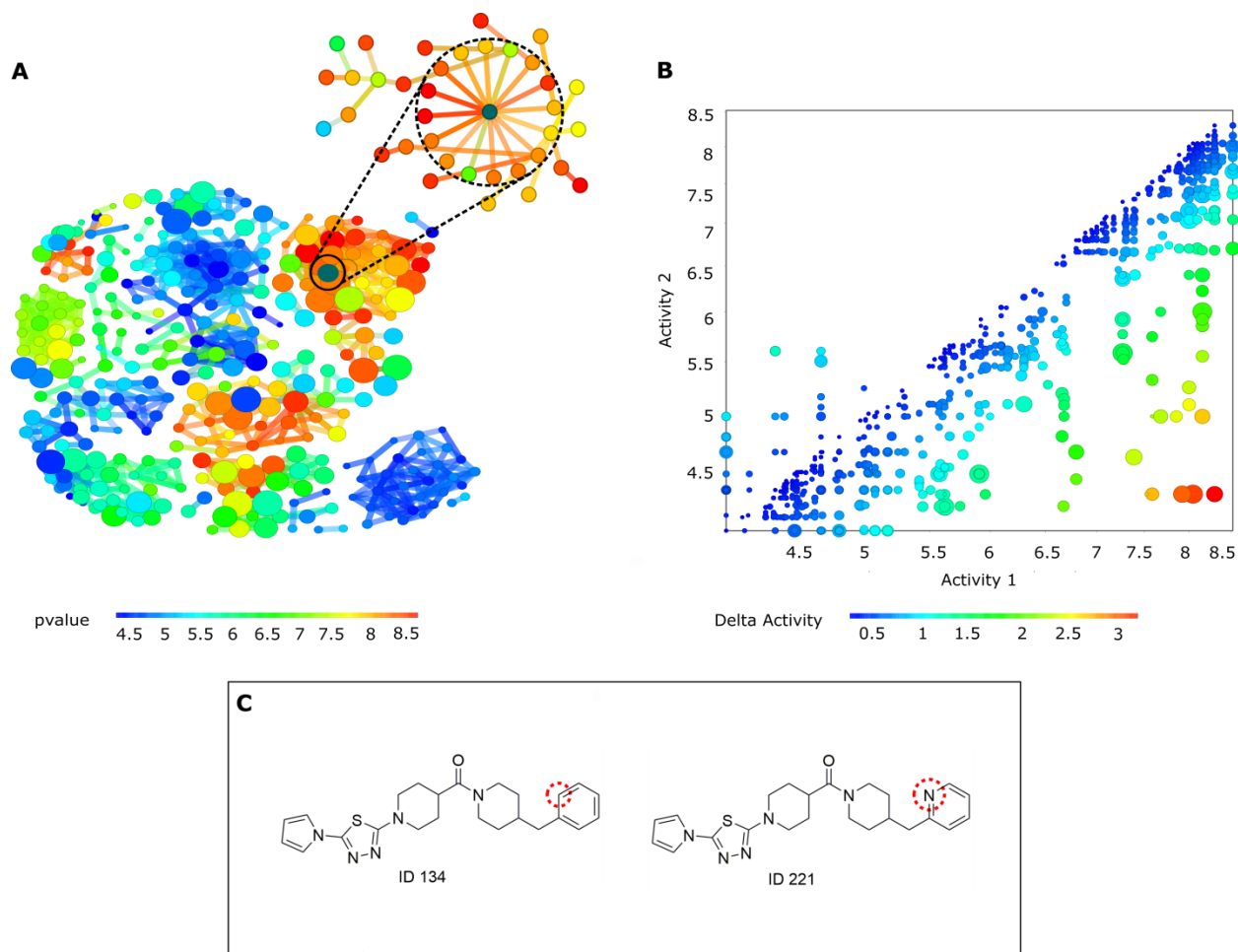


Figure S3. (A) Activity cliff set based on neighborhood similarity relationship of IC_{50} value dataset. Colors indicate p-value, where red color represents high value and blue color represents low values. The size indicates the SALI p-value of SkeleSpheres. (B) SALI plot of compound pairs. X and Y-axis represent activity values being plotted; color indicates the delta activity; higher and lower values indicated by red and blue colors, respectively. The size of the scatter indicates the SALI value. (C) Example of the compound pair (ID: 134 and ID: 221) that represents the activity cliff. The major structural differences between the compounds are highlighted with red circles.

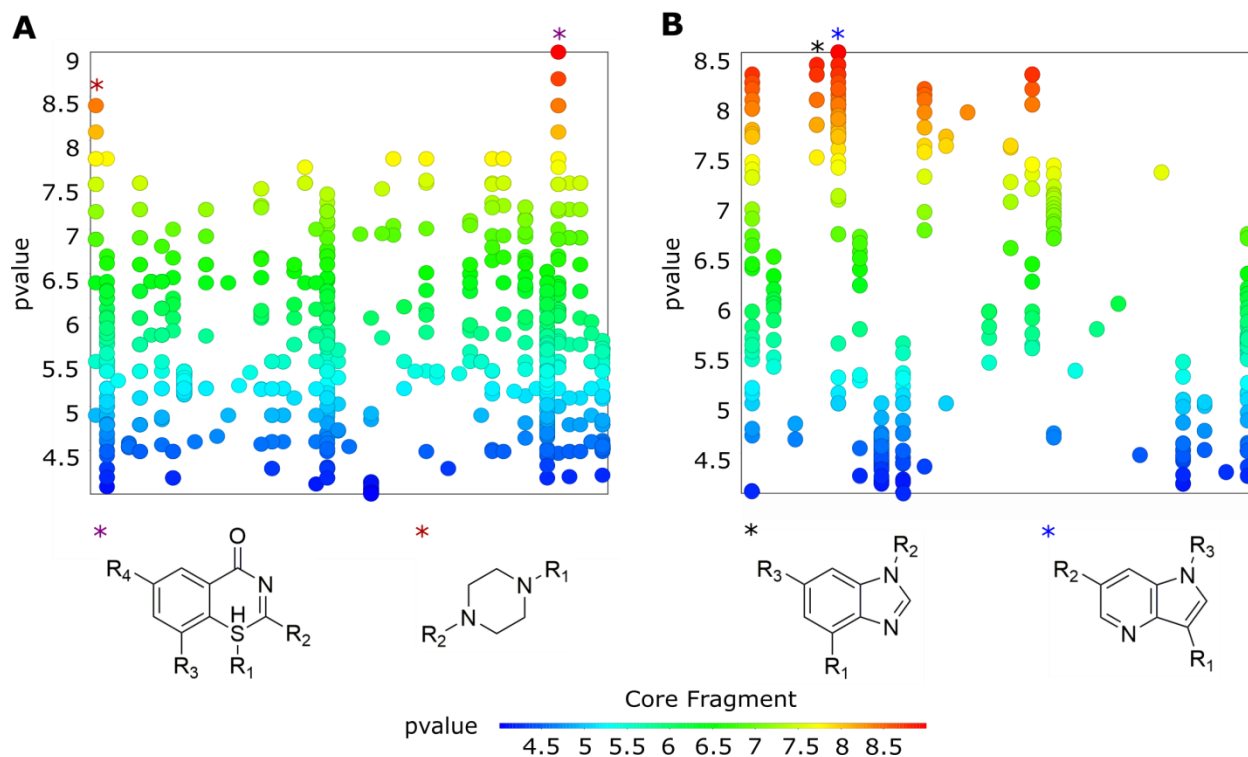


Figure S4. Core fragments vs activity (p-value) represented by scatter plot. (A) MIC value dataset. (B) IC50 values dataset. The X-axis represents the core fragments while the Y-axis represents the p-value. The color indicates the p-value, with red and blue color indicating higher and lower p-value, respectively. The asterisk mark the high biological activity fragments.

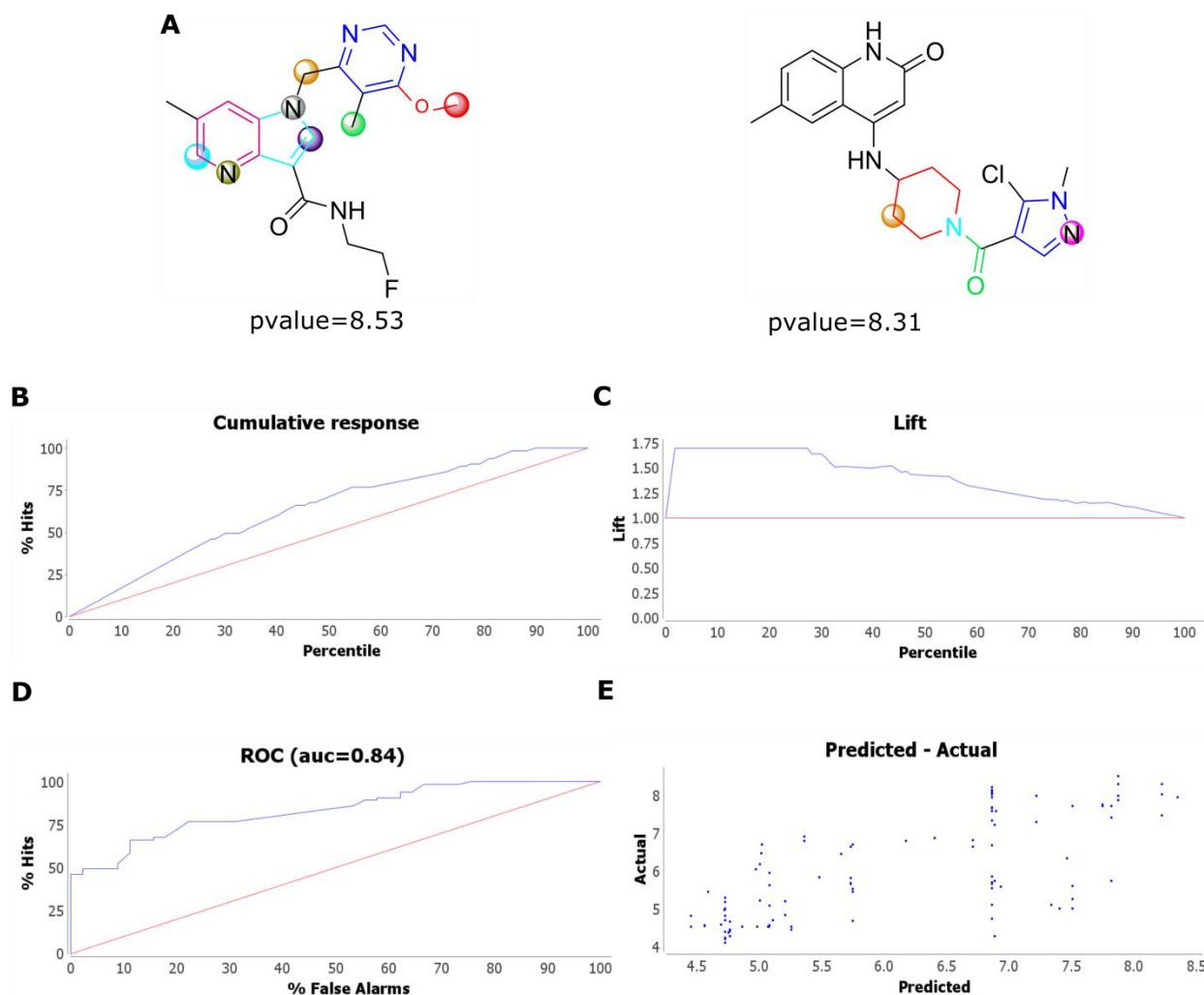


Figure S5. (A) Representative examples of small molecule inhibitors of DprE1 IC₅₀ dataset depicting the characteristic structural features contributing to SAR. (B) Cumulative response plot of the inhibitor model, showing the relation between the percentage of hits (y-axis) and percentile (x-axis). (C) Lift curve of the inhibitor model depicting observations from the percentile about the outperformance of the model over a random model. (D) ROC plot of the inhibitor representing percent of hits (y-axis) and false alarms (x-axis). (E) Predicted-actual scatter plot of the inhibitor model.

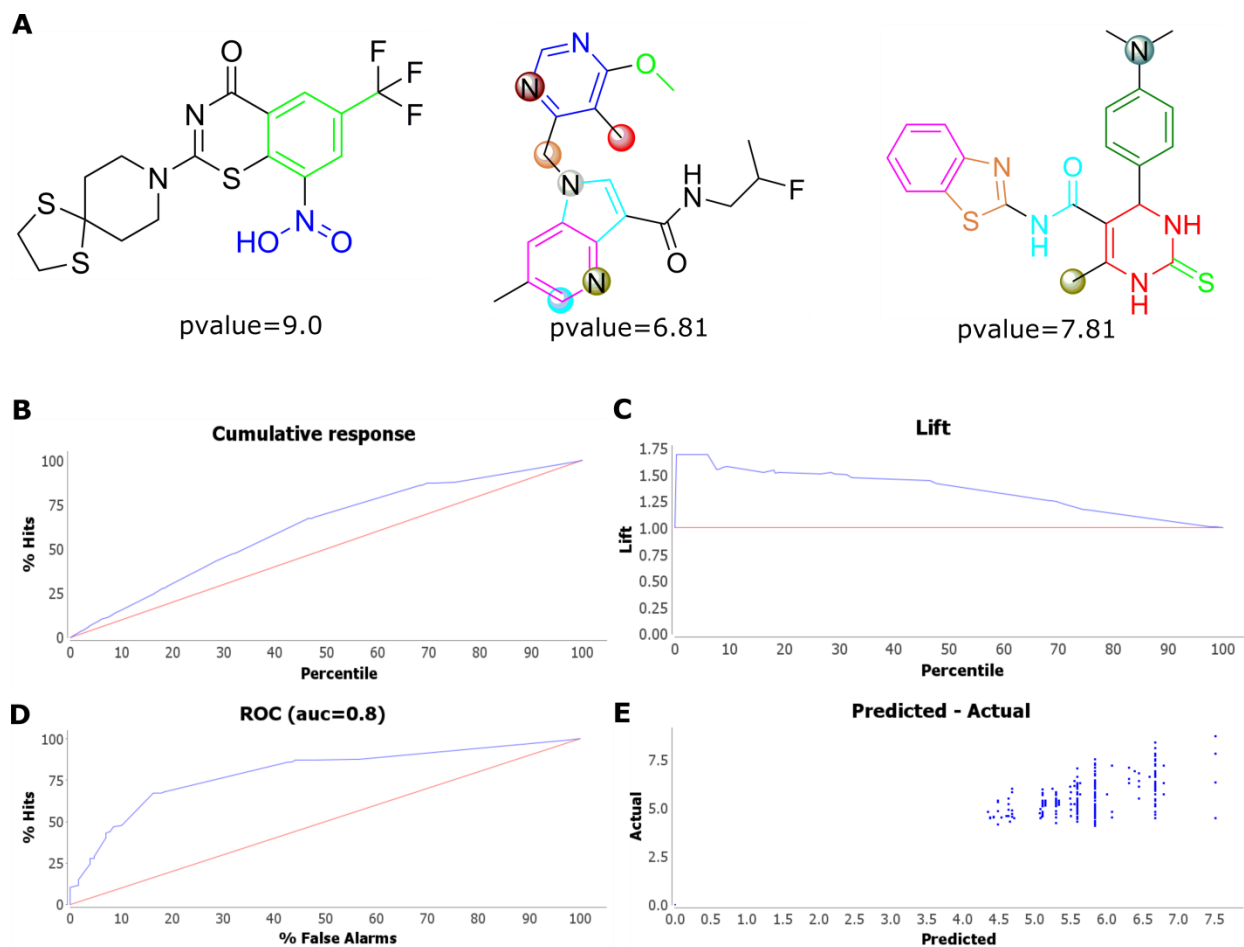


Figure S6. (A) Representative examples of small molecule inhibitors of DprE1 MIC dataset depicting the characteristic structural features contributing to SAR. (B) Cumulative response plot of the inhibitor model, showing the relation between the percentage of hits (y-axis) and percentile (x-axis). (C) Lift curve of the inhibitor model depicting observations from the percentile about the outperformance of the model over a random model. (D) ROC plot of the inhibitor representing percent of hits (y-axis) and false alarms (x-axis). (E) Predicted-actual scatter plot of the inhibitor model.