Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information

Data-driven Generation of Perturbation Networks for Relative Binding Free Energy Calculations

Jenke Scheen,[†] Mark Mackey,[‡] and Julien Michel[*,†]

†*EaStCHEM School of Chemistry, University of Edinburgh, David Brewster Road, Edinburgh EH9 3FJ, United Kingdom*

‡*Cresset Group, New Cambridge House, Bassingbourn Road, Litlington, Cambridgeshire, SG8 0SS, United Kingdom*

E-mail: mail@julienmichel.net

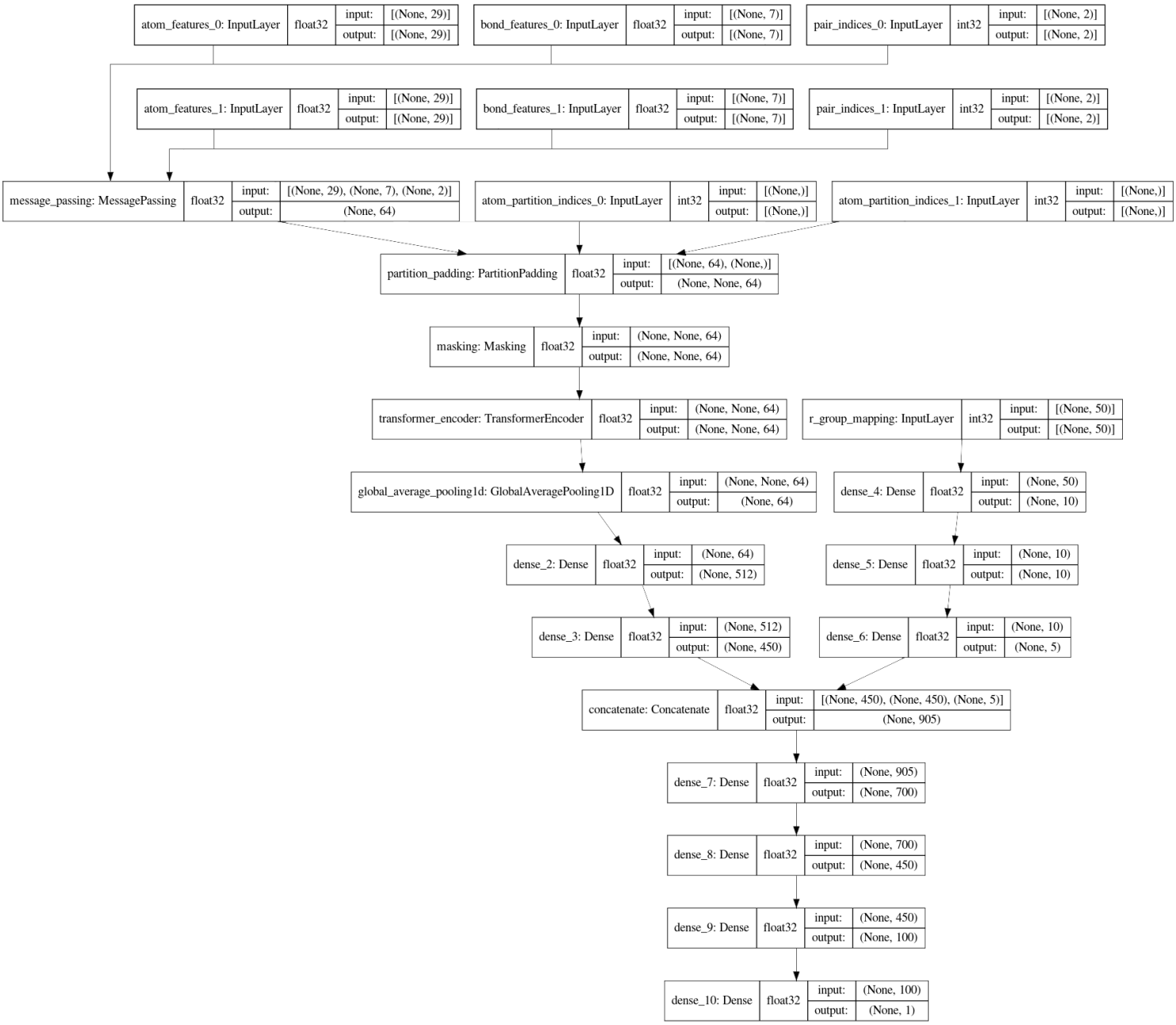

Figure S1: Low-level depiction of the 'RBFENN' siamese neural network architecture. Top to bottom: two (0 and 1) input legs are featurised into atom, bond and pair descriptors. Both legs are passed into a MessagePassing layer, which together with atom partition indices (from both legs 0 and 1) are paritioned and masked before being passed to a TransFormerEncoder layer. After a global average pooling step, two fully-connected feed-forward NN layers join with the encoded atom-mapping into a concatenation layer. Finally, three dense fully-connected feed-forward NN layers with linearly reducing numbers of parameters lead to a single-neuron layer. All dense layers in the network use ReLu activation functions except for the last single linear neuron. Each layer block depicted in this figure shows the indexed layer name (as used within TensorFlow), the class name, the dtype handled as well as the input and output dimensions.

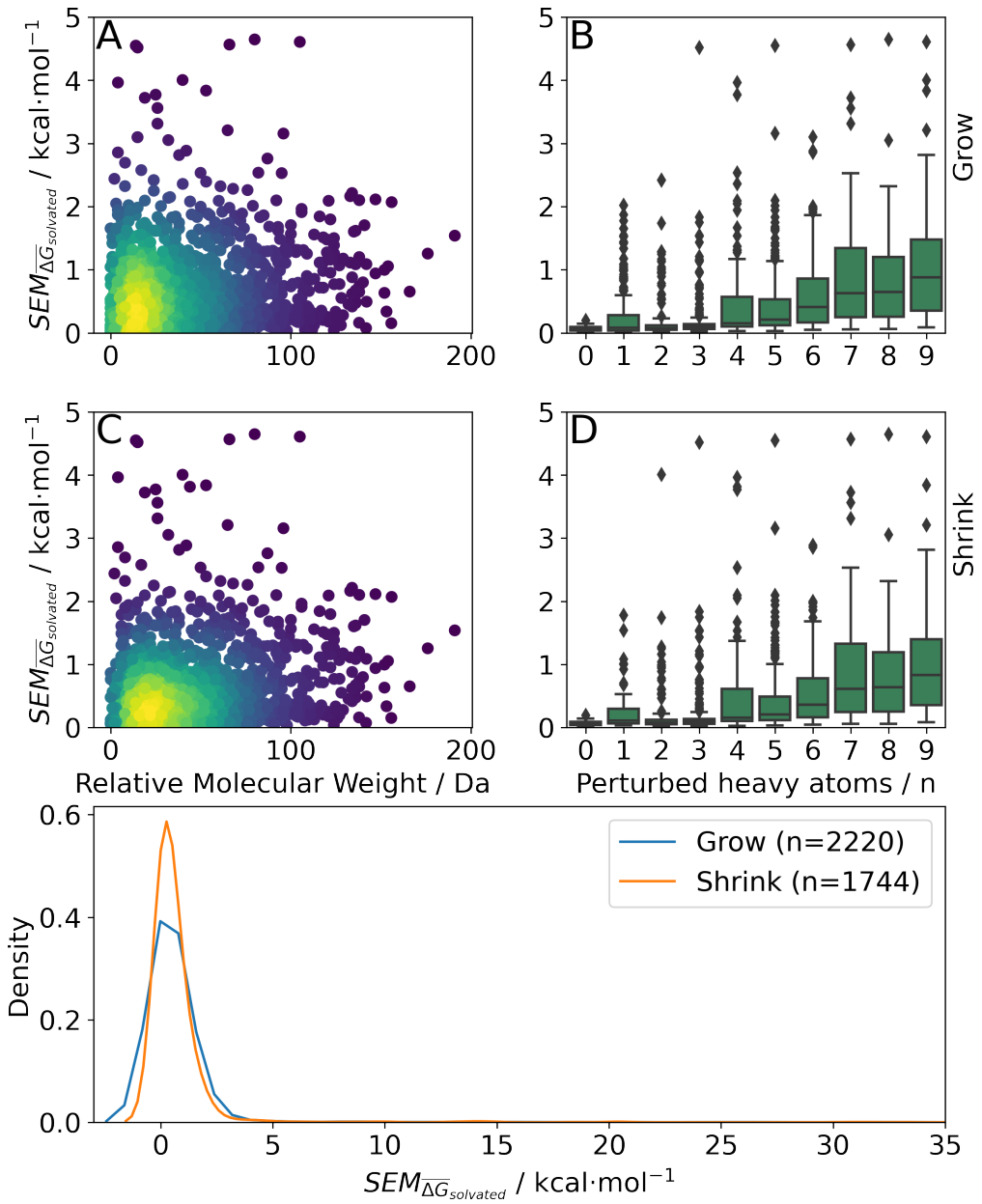

Figure S2: Summary of RBFE-Space generated using 3964 molecular perturbations grafted onto a common benzene scaffold (figure 2 main text), split by whether the perturbation involves addition ('Grow') or removal ('Shrink') of atoms. The top row shows 'Grow' perturbations, the middle shows 'Shrink' perturbations. A/C: scatterplot showing the relation between the change in molecular weight per perturbation in Da and the RBFE-Space SEM for each perturbation; colouring shows density (increasing as blue→green→yellow). B/D: boxplots of SEM $\Delta G_{solvated}$ per perturbation binned by the number of heavy atoms perturbed; horizontal lines in boxes show median values and black diamonds show outliers (95 CI). E: density plots that show the distributions of RBFE-Space SEM values for both 'Grow' (blue) and 'Shrink' (orange) type perturbations.

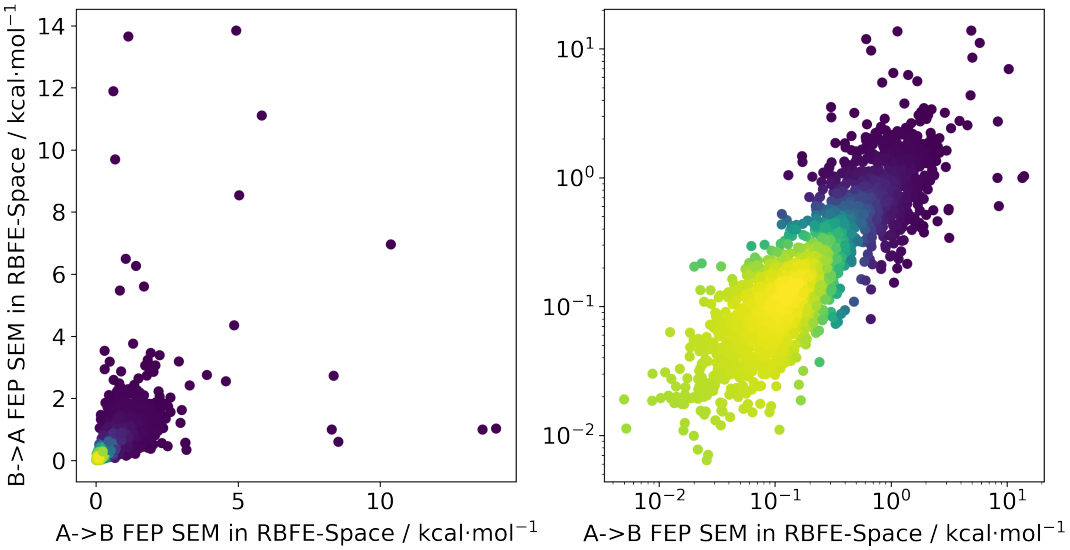

Figure S3: Comparisons of standard error of the mean (SEM) of the relative hydration free energy for all ligand pairs in RBFE-Space (n∼4000) between the two directions of a given bidirectional transformation, transforming from A→B (X axes) and back from B→A (Y axes). Shown are data on a linear scale (left-hand side) and on a logarithmic scale (right-hand side). Colour density shows the increase in data density as blue →yellow.

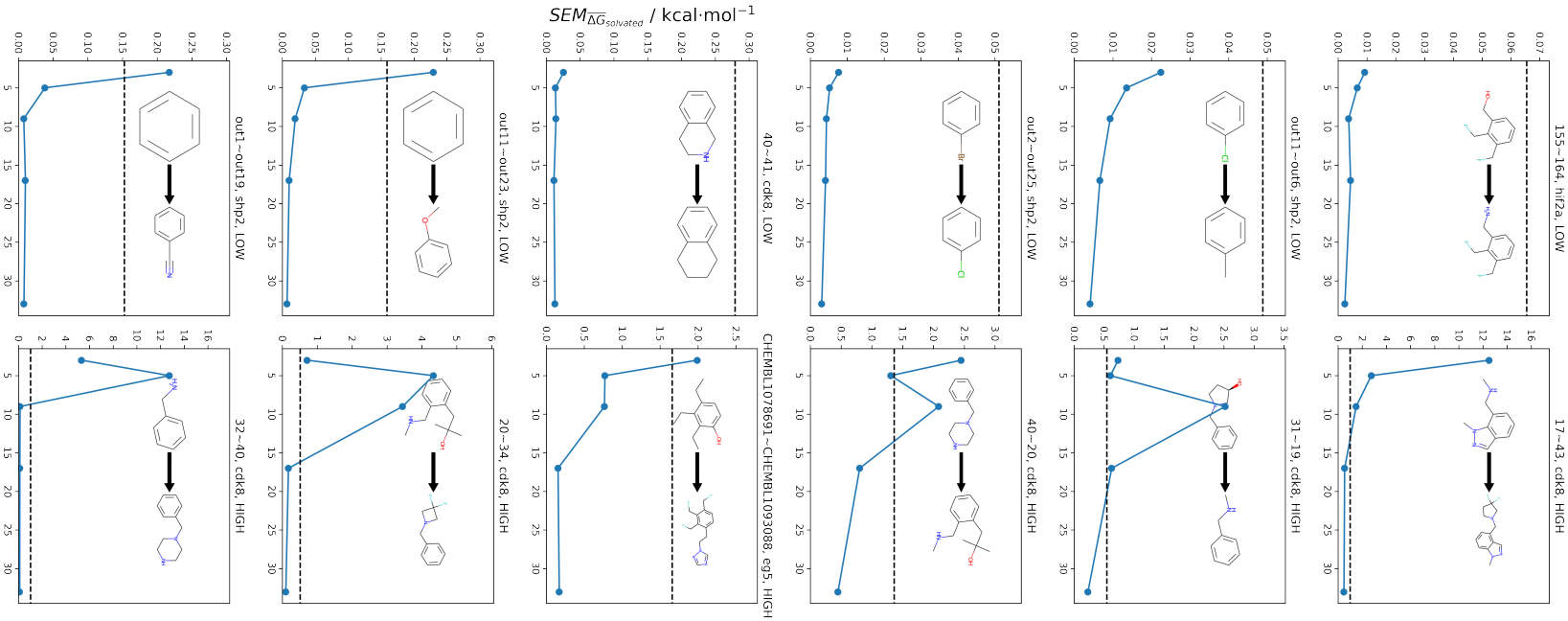

Figure S4: Molecular transformations and their statistical fluctuation represented by standard error of the mean (SEM) across five replicates shown at different numbers of λ windows. The title of each plot shows the perturbation name (the tilde signifies a transformation), the protein target and whether the expected statistical fluctuation is LOW or HIGH. The horizontal dashed line in each plot is the $\widehat{SEM}$ value as predicted by the RBFENN described in this work. Reported SEM values are $SEM_{\overline{\Delta G}_{solvated}}$ values in kcal·mol^{-1} extracted from the simulations run for the generation of the RBFE-Space training domain.

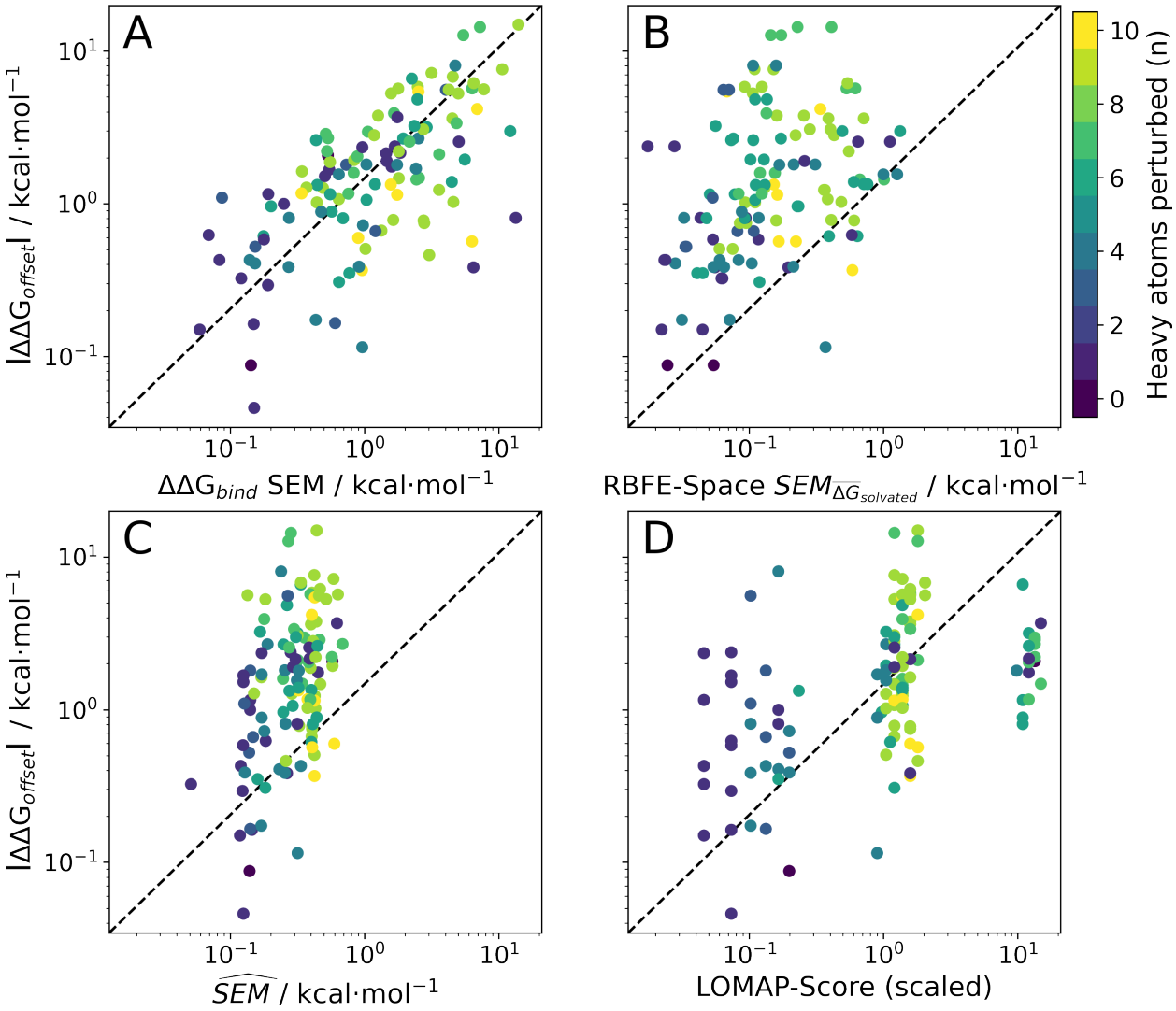

Figure S5: Scatter plots of various heuristics versus the $|\Delta\Delta G_{offset}|$ for all possible edges in the TYK2 RBFE benchmarking series (n=240). **A**: $\Delta\Delta G_{bind}$ SEM **B**: RBFE-Space SEM **C**: ML-predicted $\widehat{SEM}$ **D**: scaled LOMAP-Score (see methods section in main text body). For RBFE-Space SEM values only transformations included in RBFE-Space were included (n=124; see main text body). The colourbar shows the increase in the number of heavy atoms perturbed per perturbation in the scatter plots. See table 2 (main text body) for statistical analyses corresponding to these array comparisons.

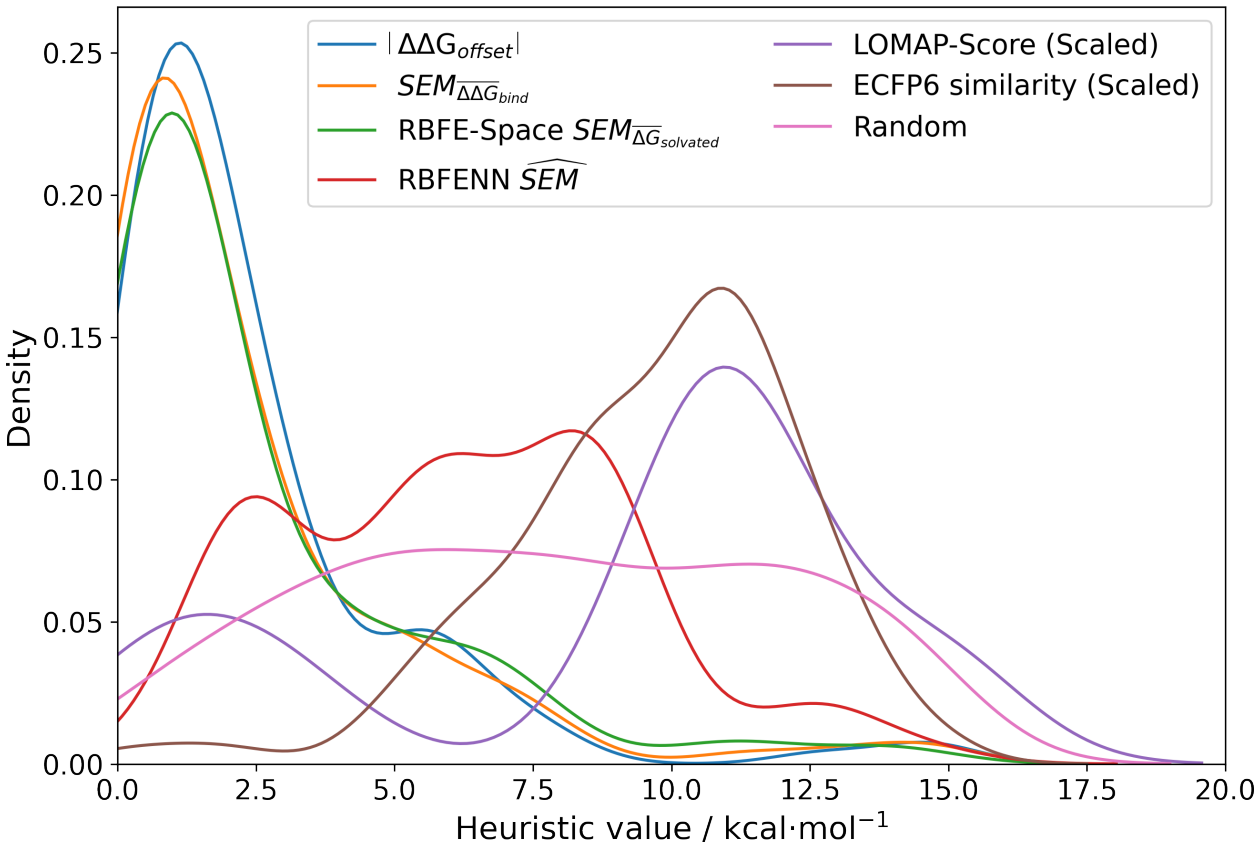

Figure S6: Array distributions using a kernel density estimation. Shown are the estimated densities of a number of heuristics that represent statistical fluctuations in RBFE transformations on a fully-connected network of the TYK2 RBFE benchmarking series. Each dataset contains 240 transformations except for RBFE-Space $SEM_{\overline{\Delta G}_{solvated}}$ which contains 124.

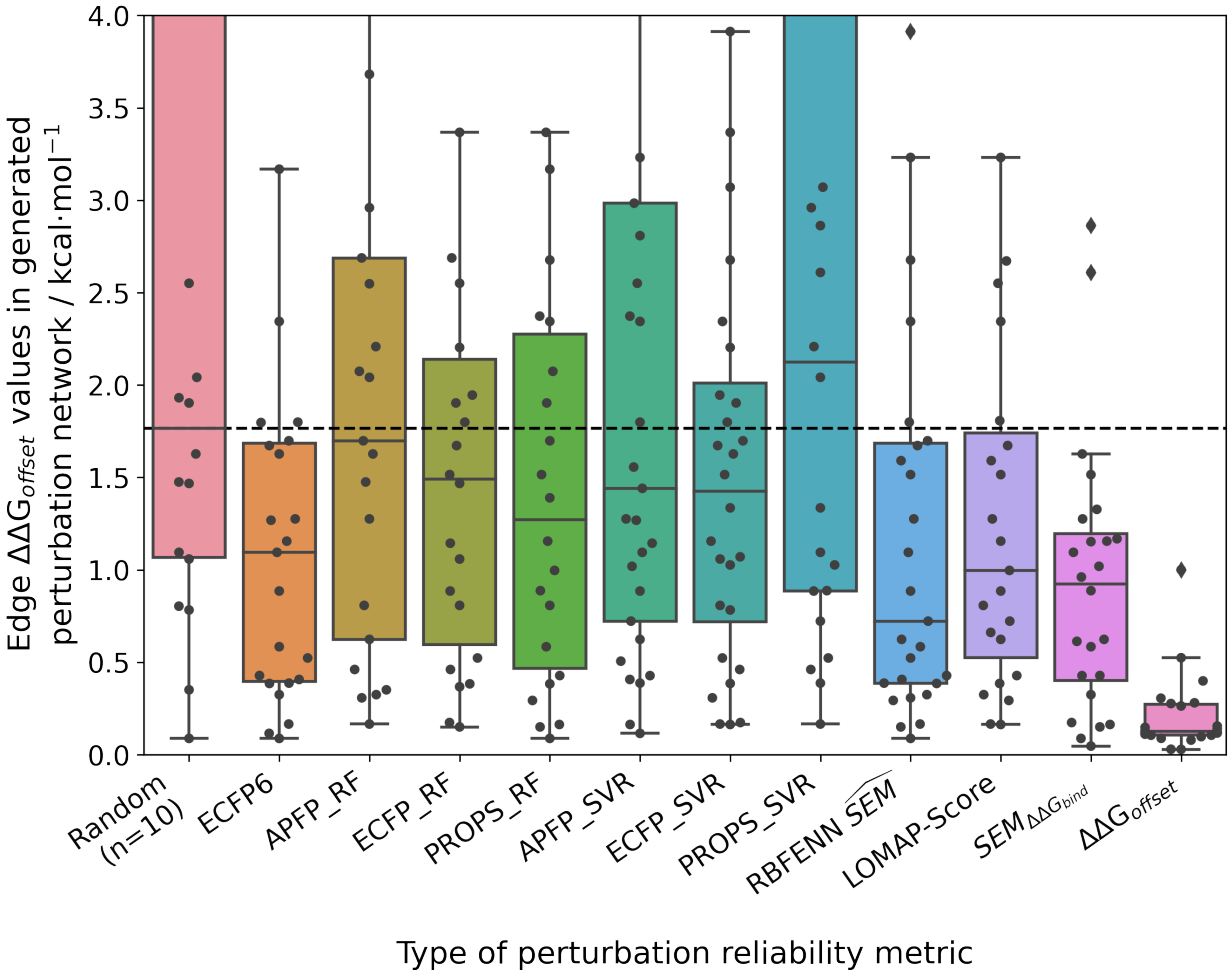

Figure S7: Boxplots depicting the distribution of $|\Delta\Delta G_{offset}|$ of edges that constituted the RBFE networks generated by various input metrics to LOMAP. The Random input metric was repeated ten times to ensure sampling of a diverse set of networks was achieved. ECFP6 is the ECFP6 tanimoto similarity between the original (i.e. with original scaffold) ligands. For RBFENN $\widehat{SEM}$, $SEM_{\Delta\Delta G_{bind}}$ and $|\Delta\Delta G_{offset}|$ the input values were scaled to an inverse 0-1 range to fit the LOMAP algorithm. The horizontal dashed line denotes the median $|\Delta\Delta G_{offset}|$ value of the Random networks.

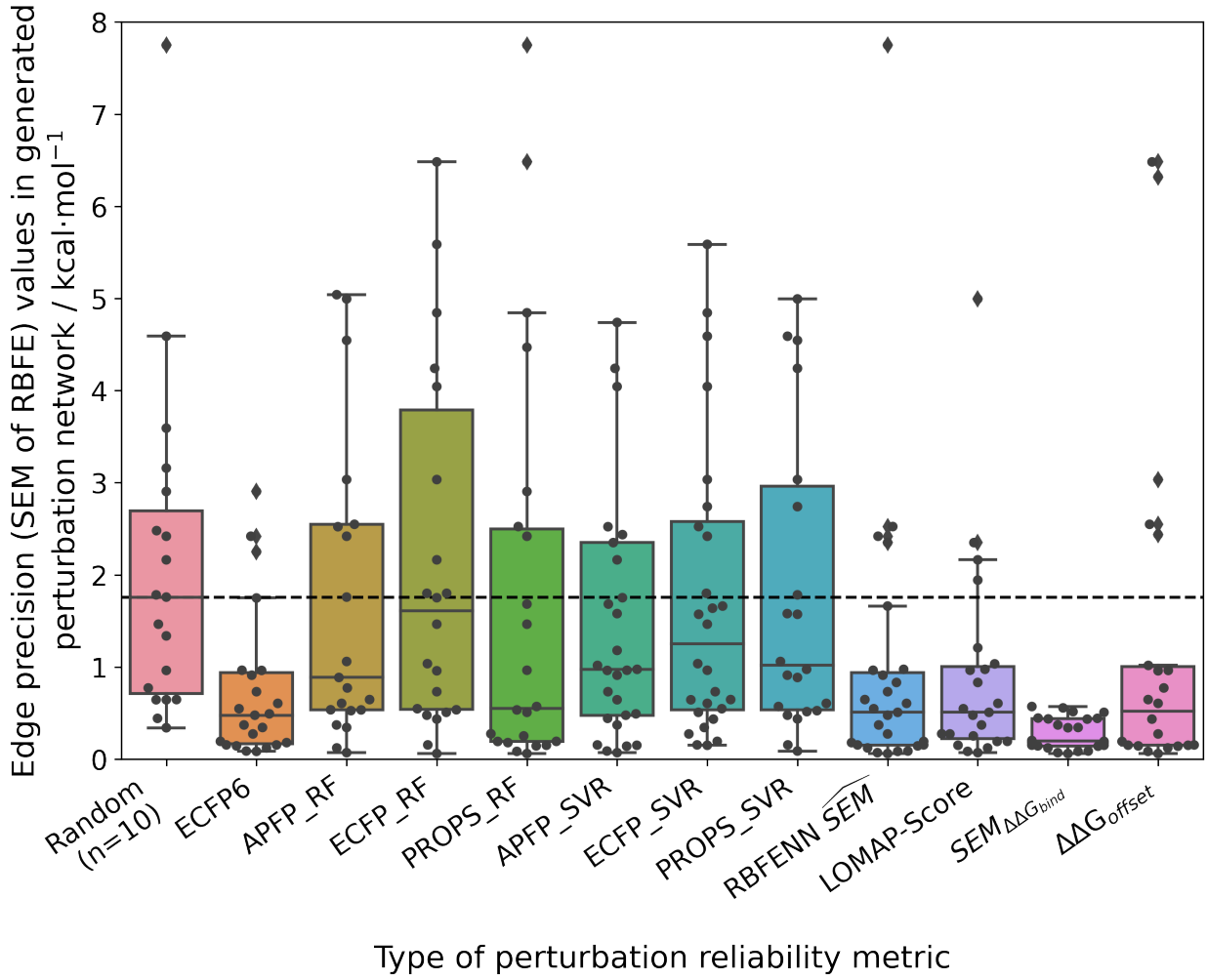

Figure S8: Boxplots depicting the distribution of SEM values of edges that constituted the RBFE networks generated by various input metrics to LOMAP. SEM values are collected from a quintuplicate FEP run on these networks. The Random input metric was repeated ten times to ensure sampling of a diverse set of networks was achieved. ECFP6 is the ECFP6 tanimoto similarity between the original (i.e. with original scaffold) ligands. For RBFENN $\widehat{SEM}$, $SEM_{\Delta\Delta G_{bind}}$ and $|\Delta\Delta G_{offset}|$ the input values were scaled to an inverse 0-1 range to fit the LOMAP algorithm. The horizontal dashed line denotes the median SEM value of the Random networks.

TYK2 - Random Forest (Molecular properties)

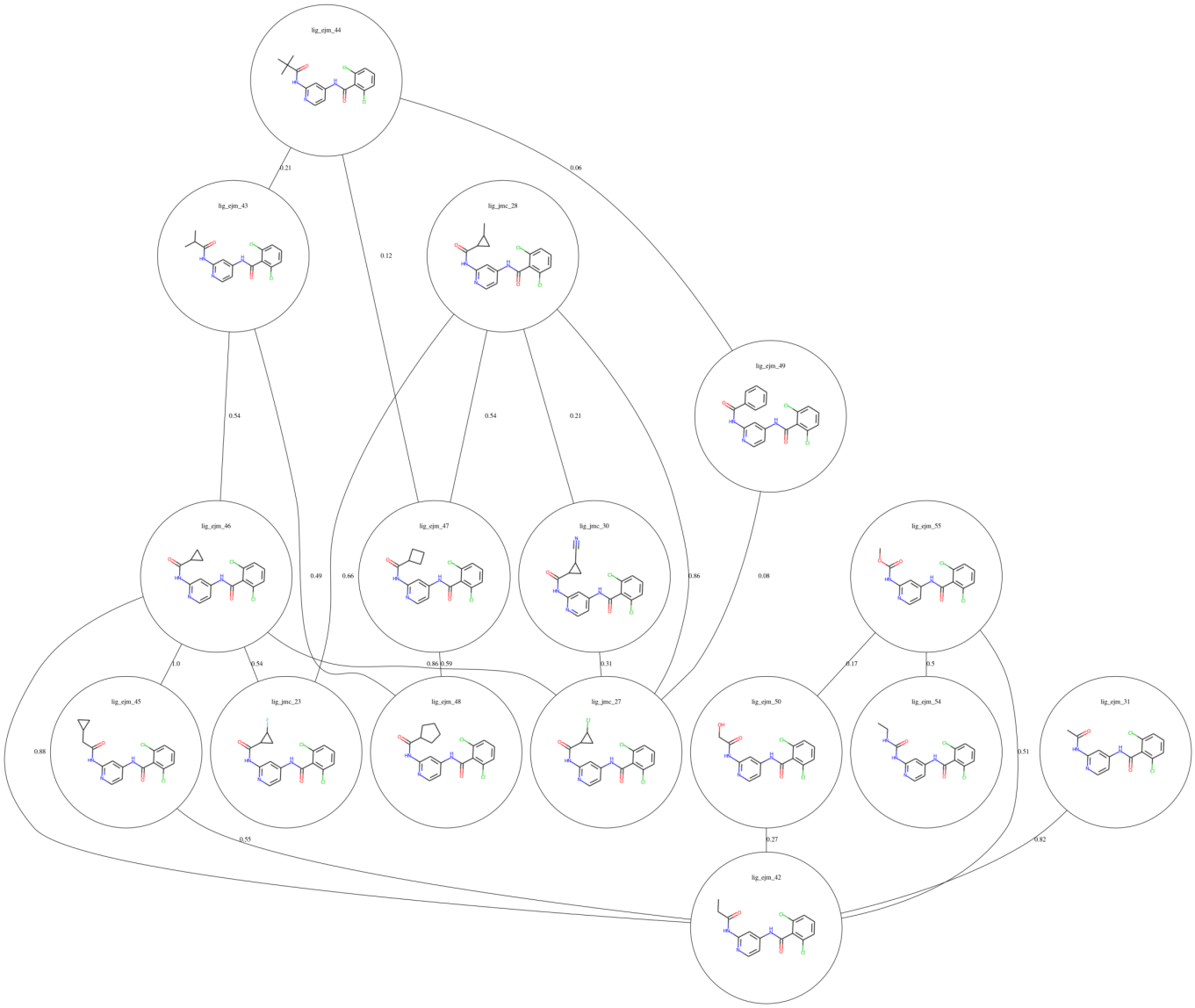

Figure S9: The TYK2 perturbation network as suggested by LOMAP using $\widehat{SEM}$ predicted by a random forest using molecular descriptors as input. Each node in the network contains the molecular structure and the ligand name; each edge in the network is annotated with the predicted $\widehat{SEM}$ value.

TYK2 - ECFP6

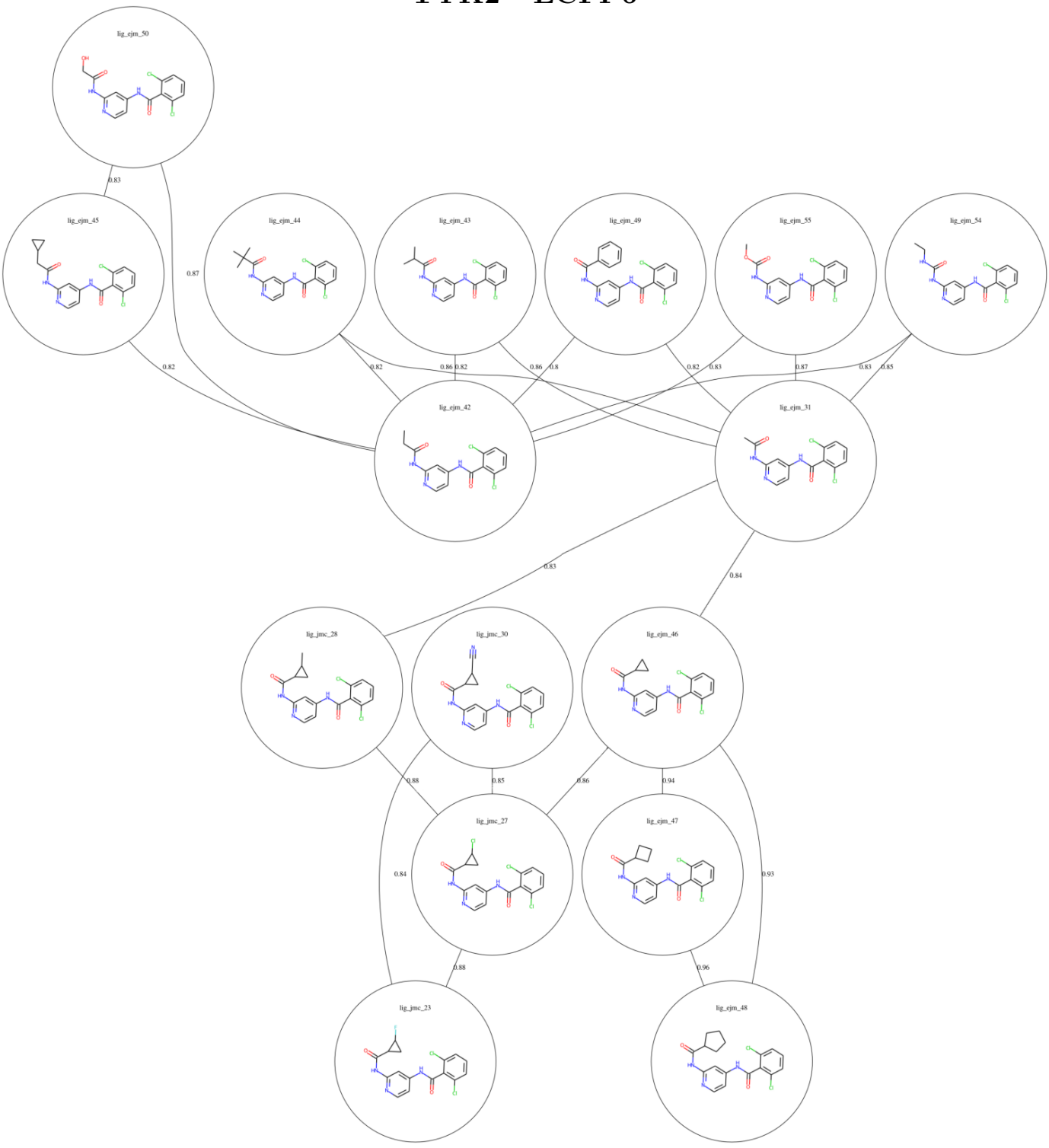

Figure S10: The TYK2 perturbation network as suggested by LOMAP using ECFP6 tanimoto similarities (on original ligands) as input. Each node in the network contains the molecular structure and the ligand name; each edge in the network is annotated with the similarity value.

TYK2 - RANDOM

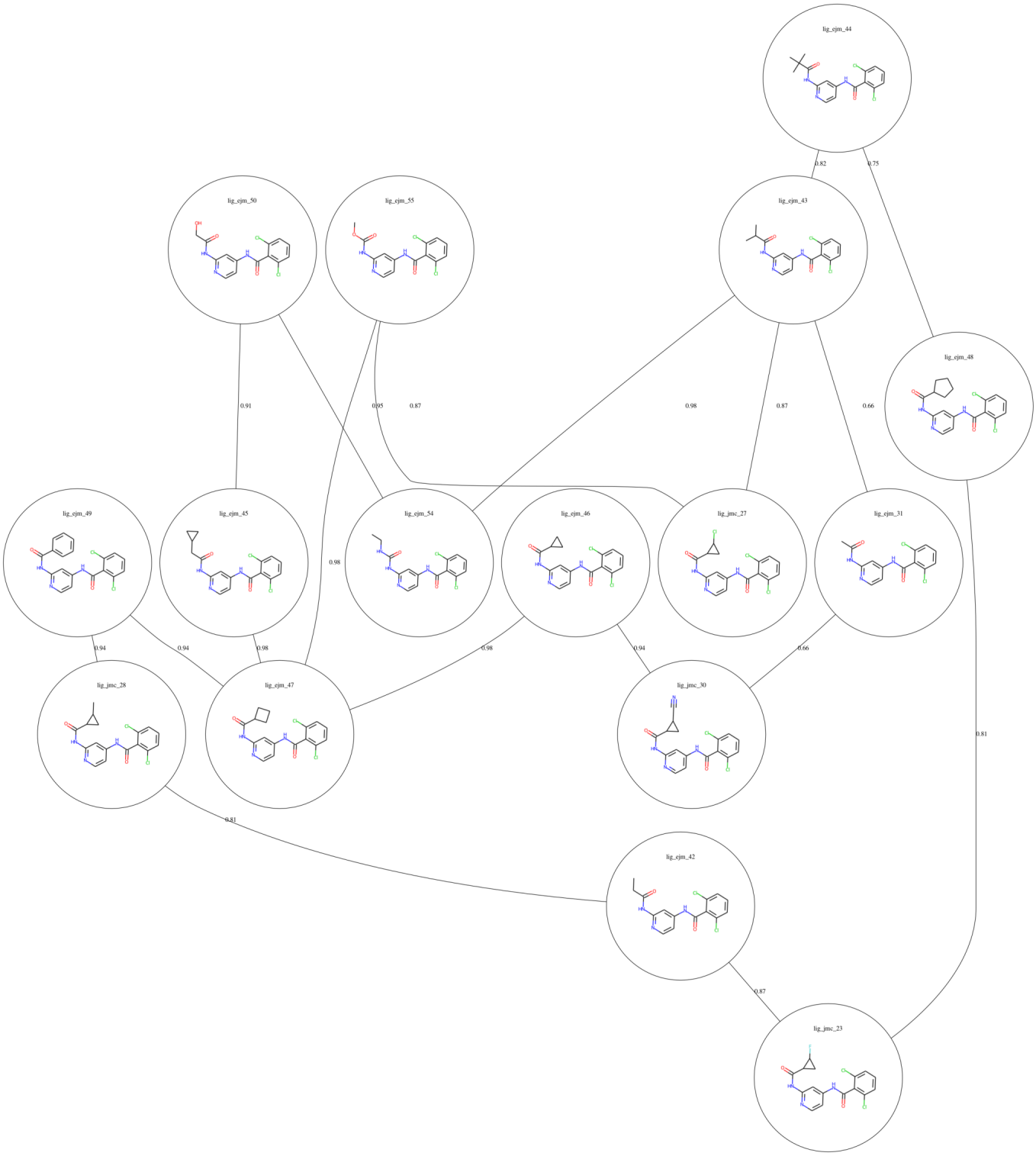

Figure S11: The TYK2 perturbation network as suggested by LOMAP using random values as input. Each node in the network contains the molecular structure and the ligand name; each edge in the network is annotated with the random value.

TYK2 - RBFENN

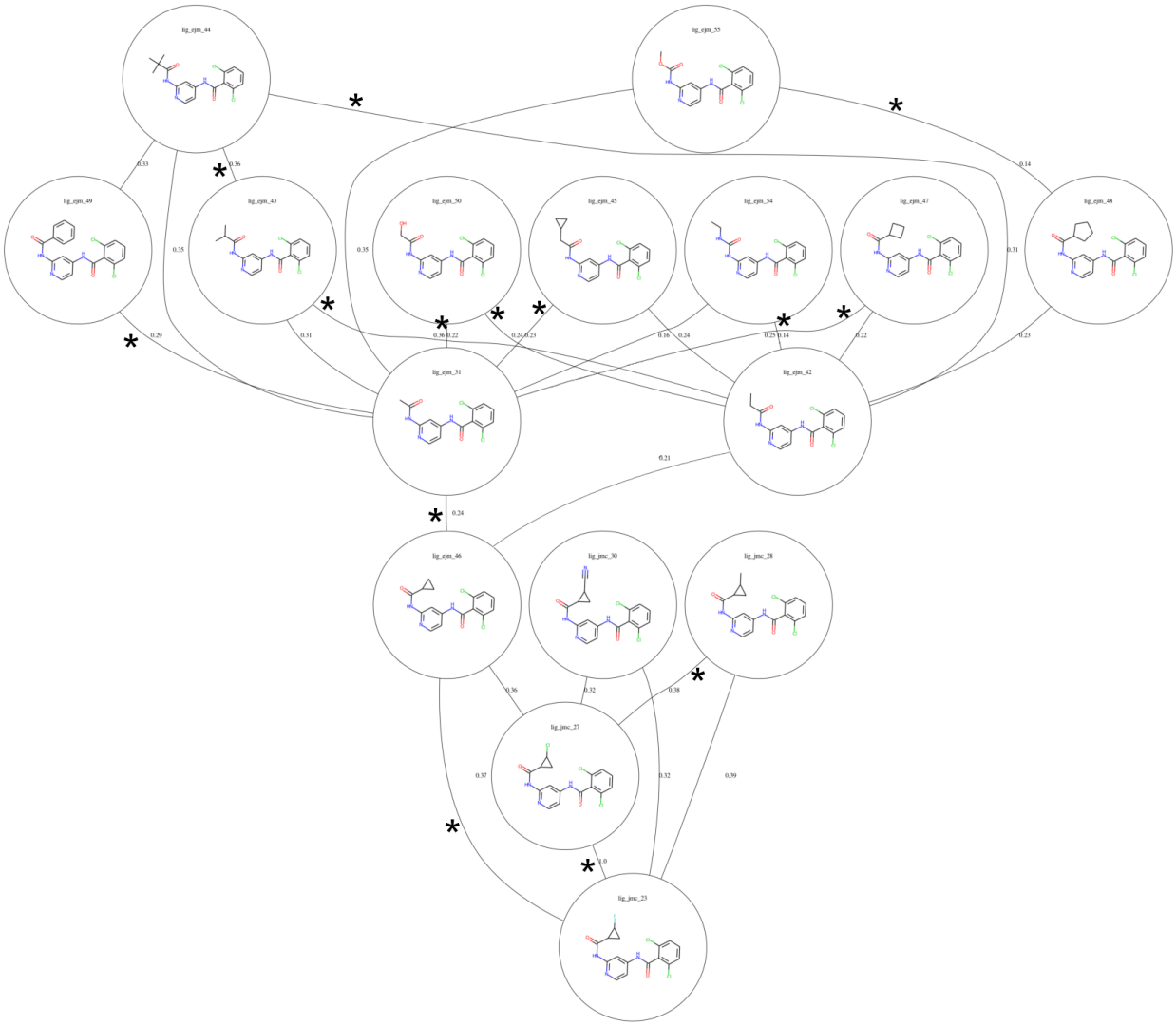

Figure S12: The TYK2 perturbation network as suggested by LOMAP using the RBFENN-predicted $\widehat{SEM}$ score as input. Each node in the network contains the molecular structure and the ligand name; each edge in the network is annotated with the RBFENN-predicted $\widehat{SEM}$ value that has been scaled to [0-1] to allow proper handling by the LOMAP algorithm. Asterisks (*) indicate edges that are shared between the RBFENN and LOMAP networks.

TYK2 - LOMAP-Score

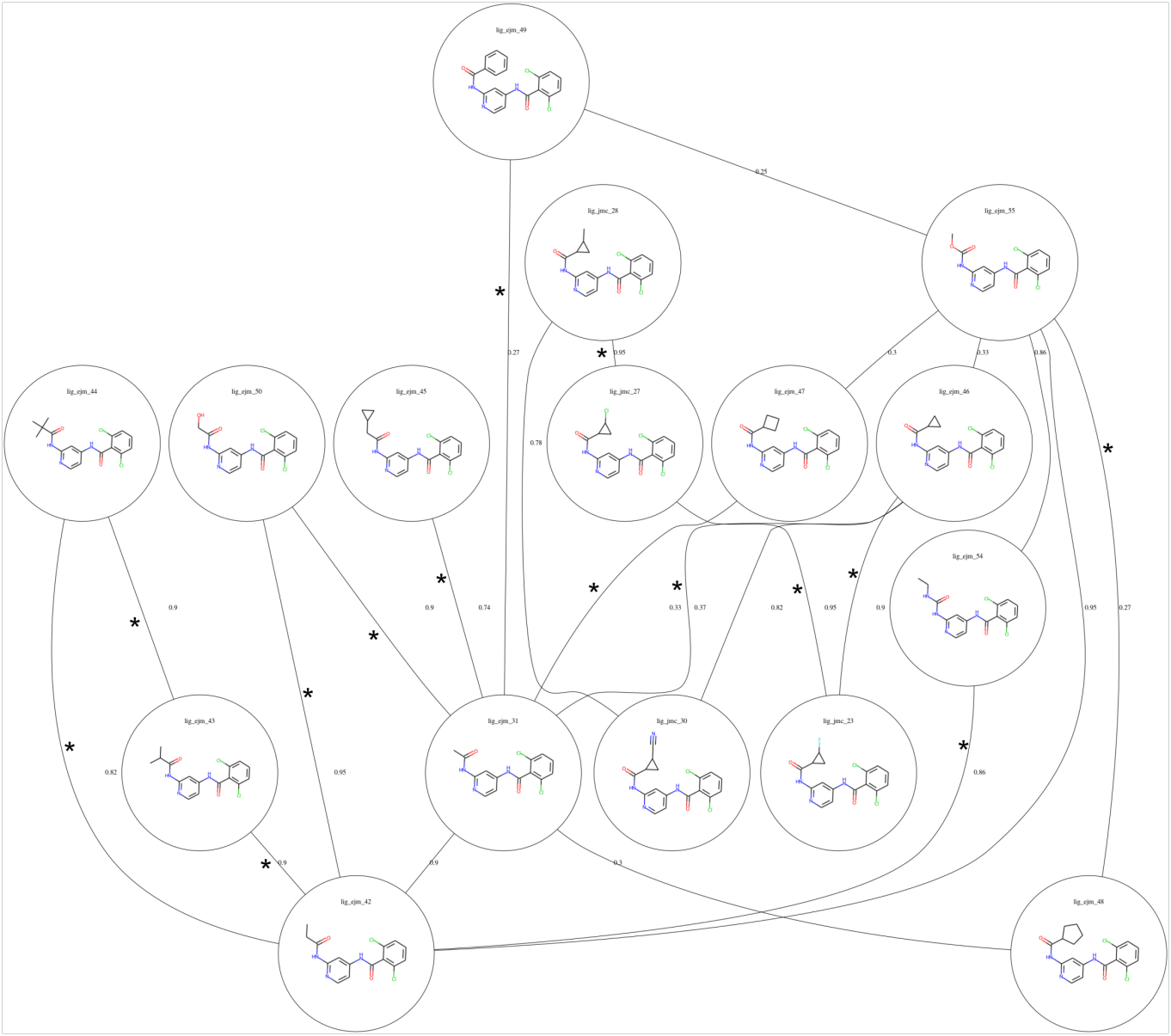

Figure S13: The TYK2 perturbation network as suggested by LOMAP using the LOMAP-Score as input. Each node in the network contains the molecular structure and the ligand name; each edge in the network is annotated with the assigned LOMAP-Score value. Asterisks (*) indicate edges that are shared between the RBFENN and LOMAP networks.

TYK2 - $|\Delta\Delta G_{offset}|$

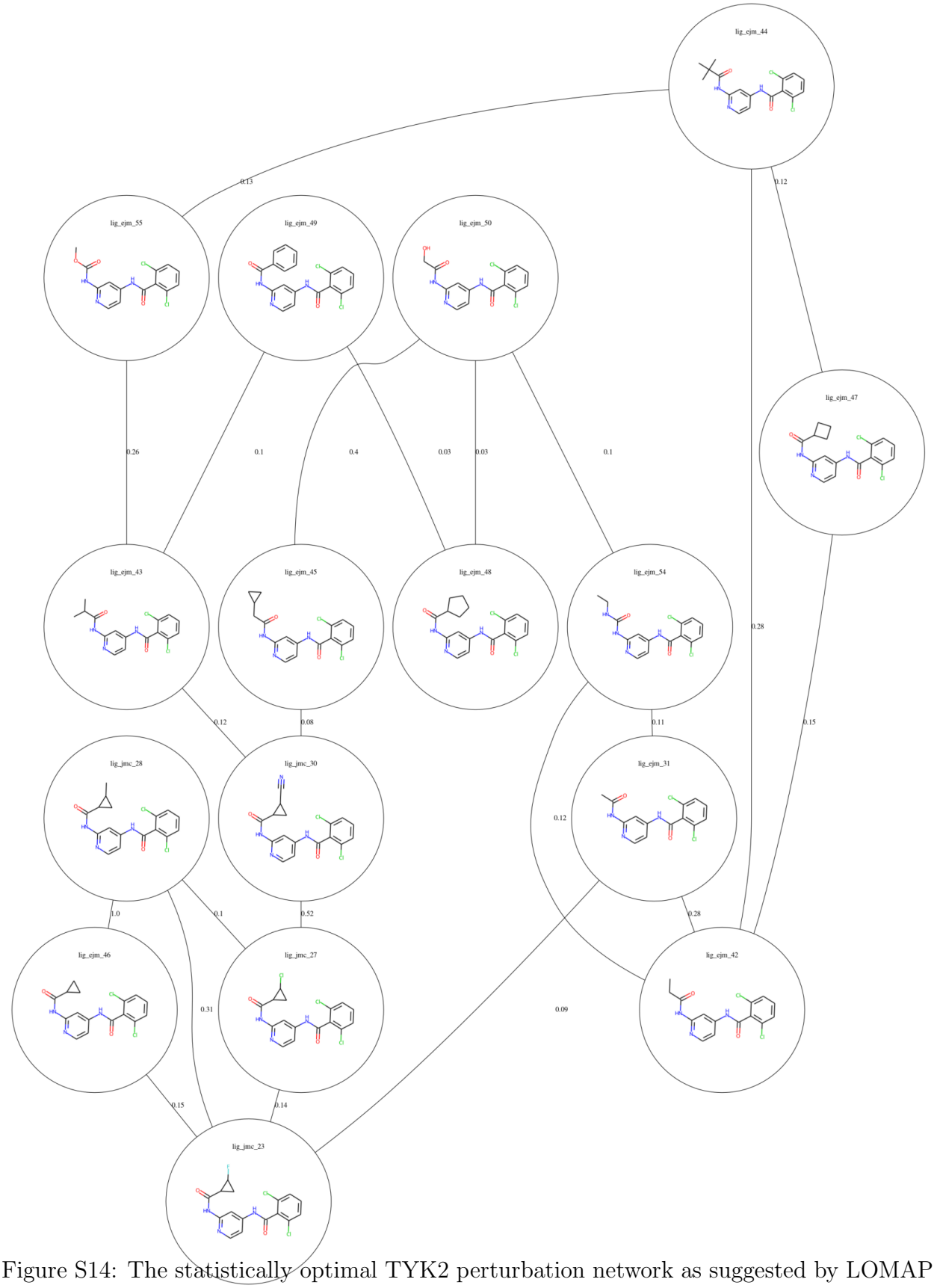

Figure S14: The statistically optimal TYK2 perturbation network as suggested by LOMAP using $|\Delta\Delta G_{offset}|$ values as input. Each node in the network contains the molecular structure and the ligand name; each edge in the network is annotated with the $|\Delta\Delta G_{offset}|$ value that has been scaled to [0-1] to allow proper handling by the LOMAP algorithm.

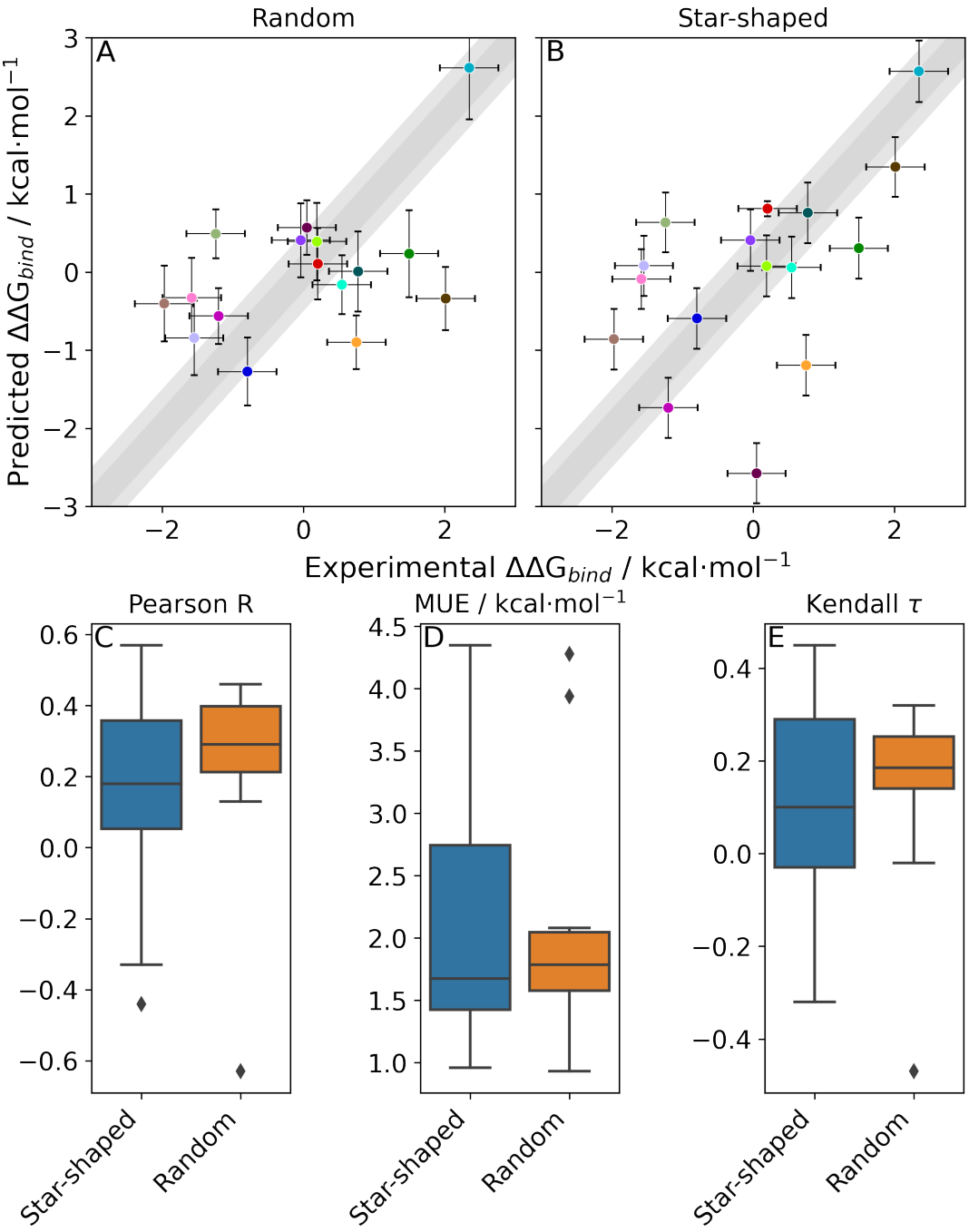

Figure S15: Comparison of predictive performances for TYK2 of perturbation networks generated using random selection of edges and the star-shaped approach. **A**/**B**: scatterplots of representative (i.e. $n = 1$) random and star-shaped networks' RBFE predictions compared to experimental measures in kcal·mol^{-1}. Ligands are coloured for direct comparison of positioning between the two plots. **C-E**: boxplots showing distributions of statistical performances for the complete collection of networks for both star-shaped ($n = 16$) and random ($n = 10$) network approaches.

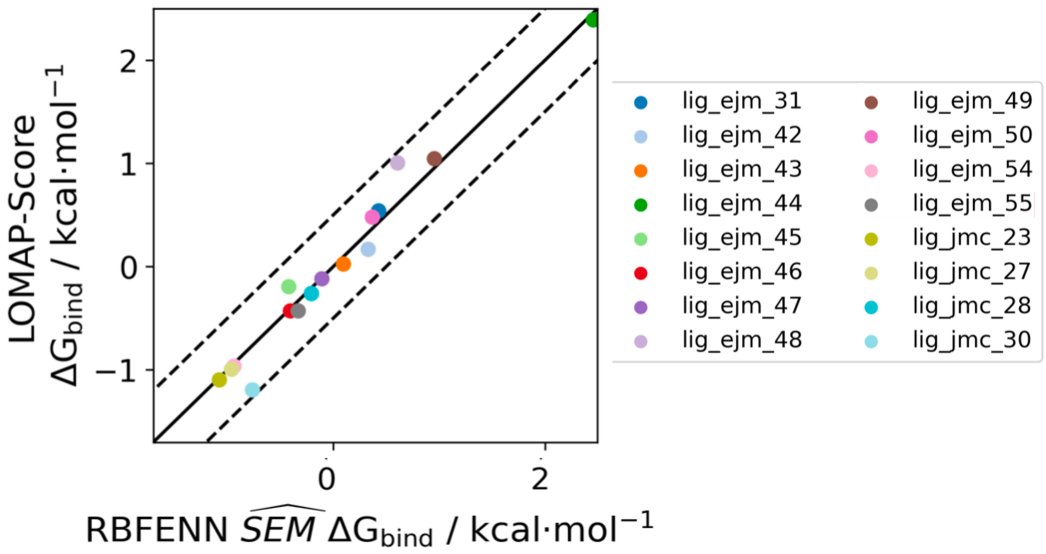

Figure S16: Scatterplots comparing RBFE predictions per ligand of the TYK2 benchmarking series using the LOMAP-Score and RBFENN derived networks. The 1 kcal·mol^{-1} error bound is shown with dashed lines.

TNKS2 - RBFENN

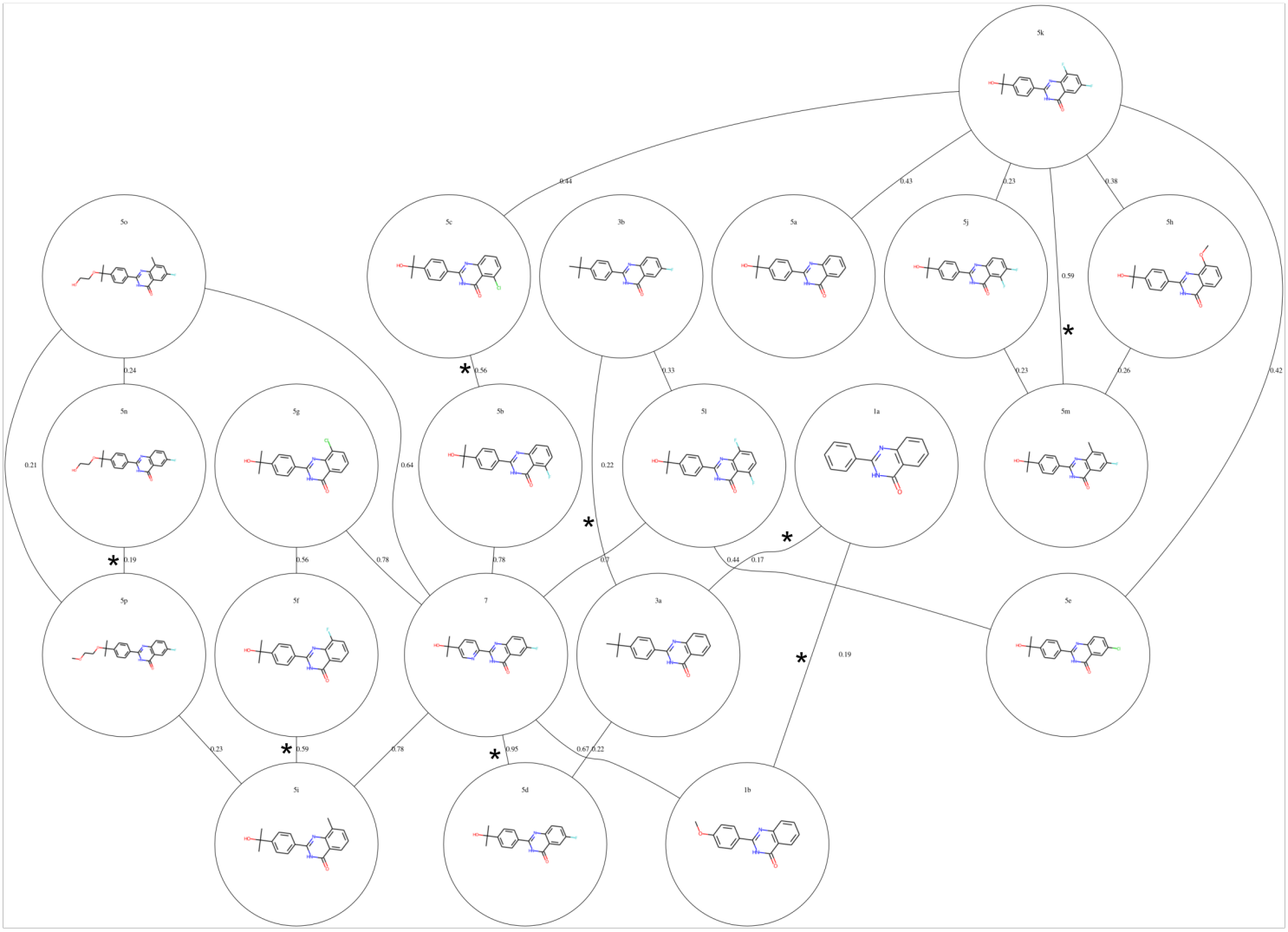

Figure S17: The TNKS2 perturbation network as suggested by LOMAP using the RBFENN-predicted $\widehat{SEM}$ score as input. Each node in the network contains the molecular structure and the ligand name; each edge in the network is annotated with the RBFENN-predicted $\widehat{SEM}$ value that has been scaled to [0-1] to allow proper handling by the LOMAP algorithm. Asterisks (*) indicate edges that are shared between the RBFENN and LOMAP networks. For this series, the six ligands with a +1 formal charge have been excluded.

TNKS2 - LOMAP-Score

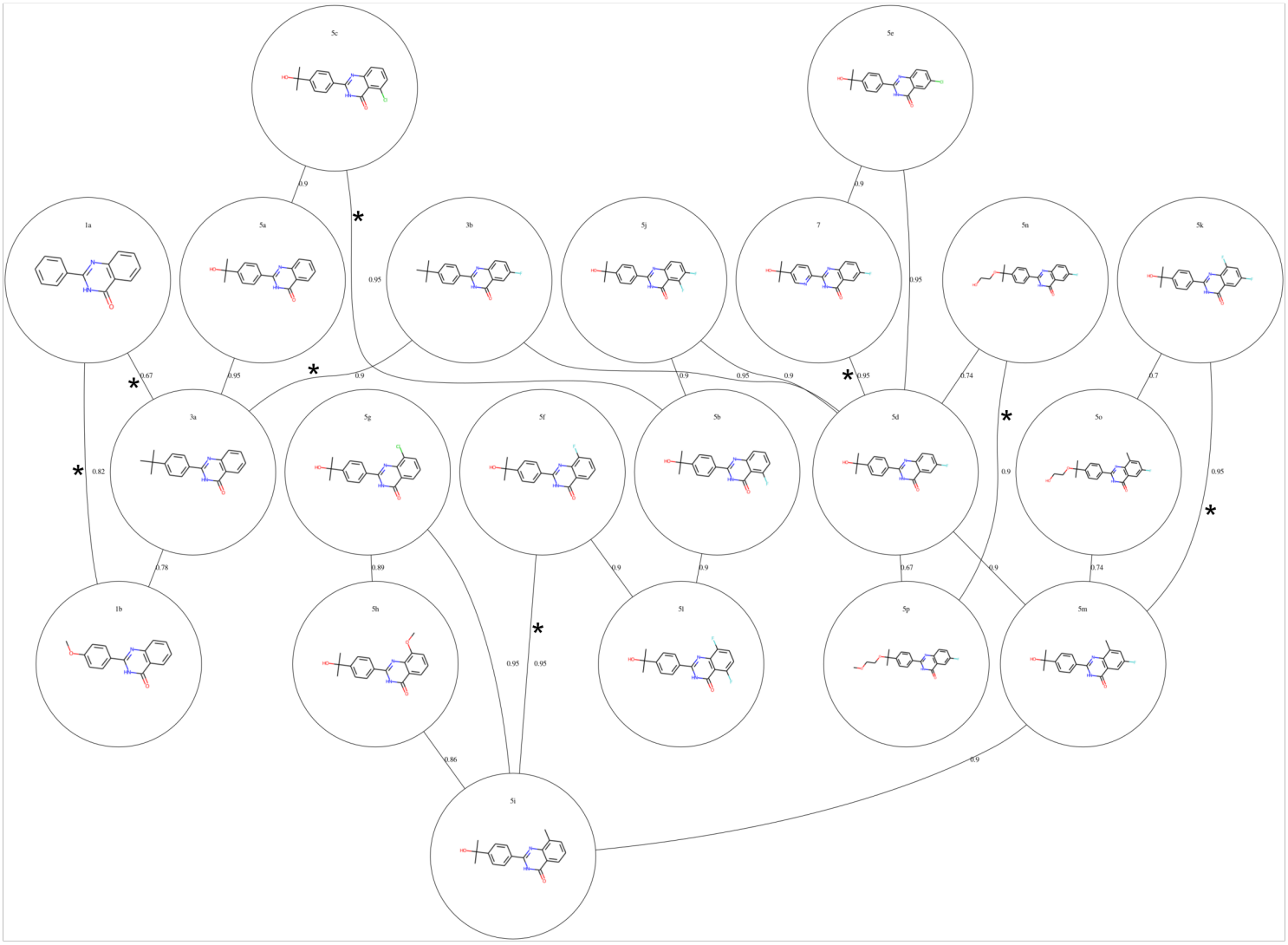

Figure S18: The TNKS2 perturbation network as suggested by LOMAP using the LOMAP-Score as input. Each node in the network contains the molecular structure and the ligand name; each edge in the network is annotated with the assigned LOMAP-Score value. Asterisks (*) indicate edges that are shared between the RBFENN and LOMAP networks. For this series, the six ligands with a +1 formal charge have been excluded.

Table S1: In-house results provided by Cresset on the neutral ligands of the TNKS2 RBFE benchmarking series. Shown are results of an RBFE run using a network with 70 edges run using Flare V4. Columns contain data on the experimental binding affinity, the experimental error, the RBFE-predicted binding affinity and the absolute error between experimental and predicted binding affinity for each ligand. Shown below the table are statistics as generated by Flare; Pearson R for this data is 0.75. See table S2 for edges and methodology.

Molecule	Experimental Activity	Error	Predicted Activity	abs(err)
1a	-8.55	0.3	-8.07	0.48
1b	-9.93	0.28	-10.04	0.11
3a	-10.99	0.22	-10.99	0
3b	-11.51	0.29	-10.83	0.68
5a	-10.76	0.23	-10.43	0.33
5b	-10.47	0.22	-11.11	0.64
5c	-9.95	0.28	-9.8	0.15
5d	-10.88	0.23	-10.3	0.58
5e	-10.1	0.46	-9.39	0.71
5f	-10.25	0.22	-11	0.75
5g	-10.8	0.3	-11.21	0.41
5h	-10.05	0.28	-9.57	0.48
5i	-12.07	0.31	-10.94	1.13
5j	-11.07	0.27	-11.53	0.46
5k	-10.96	0.28	-11.01	0.05
5l	-10.09	0.25	-11.47	1.38
5m	-12.68	0.33	-11.06	1.62
5n	-10.7	0.45	-10.54	0.16
5o	-12.03	0.69	-13.75	1.72
5p	-10.5	0.29	-11.02	0.52
7	-8.39	0.76	-8.65	0.26

Pearson r^2: 0.56 (95%CI 0.19-0.81)
MUE: 0.60 (95%CI 0.41-0.81) kcal·mol^{-1}

Table S2: perturbations run in-house by Cresset on TNKS2 (see table S1). Shown are relative binding free energy predictions for each edge in the chosen RBFE network ($n = 70$) in kcal·mol^{-1} for both the forward (A→B) and reverse (B→A) transformation. This RBFE campaign was run using Flare V4 with a total of 754 λ windows.

Edge	A→B	B→A	Edge	A→B	B→A
1a∼1b	-2.15	2.14	5d∼5m	-0.25	0.5
1a∼3a	-3.06	3.39	5d∼5n	0.54	1.95
1b∼3a	-1.1	1.03	5d∼5o	-8.06	6.21
1b∼3b	-1.12	0.49	5d∼5p	-1.02	1.19
3a∼3b	0.24	-0.37	5d∼7	1.26	-2.06
3a∼5a	0.53	-0.5	5e∼7	0.87	-0.08
3a∼5b	-0.19	0.35	5f∼5g	-0.27	0.34
3a∼5f	-0.06	0.25	5f∼5h	1.61	-1.84
3b∼5d	0.45	-0.54	5f∼5i	0.24	-0.04
5a∼5b	-0.71	0.68	5f∼5l	-0.51	0.82
5a∼5d	0.07	-0.21	5g∼5h	1.85	-2.08
5a∼5f	-0.81	0.66	5i∼5l	-0.74	0.9
5b∼5c	1.3	-1.45	5j∼5k	0.98	-0.58
5b∼5j	-0.24	0.42	5k∼5m	0.15	-0.08
5b∼5l	-0.33	0.46	5m∼5o	-7.17	4.29
5c∼5l	-1.65	1.72	5n∼5p	-0.76	0.96
5d∼5e	0.88	-1.21			
5d∼5j	-1.27	1.39			
5d∼5k	-0.5	0.48			