Optimal Designs for Pairwise Calculation: an Application to Free Energy Perturbation in Minimizing Prediction Variability

Qingyi Yang* Woodrow Burchett† Gregory S. Steeno† Shuai Liu‡
Mingjun Yang‡ David L. Mobley§ Xinjun Hou*

August 28, 2019

SUPPLEMENTARY INFORMATION

Mathematical Structure

The following is an illustration of sub-matrix $\mathbf{A}_{\Delta G}$ mapping experimental values ΔG_{Exp}, following the notation introduced in the main text:

$$\mathbf{A}_{\Delta G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & . & . & . & . & . & 0 \\ 0 & . & . & . & . & . & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}_{(m_{Exp},N)}$$

Illustration of sub-matrix $\mathbf{A}_{\Delta\Delta G}$ mapping pairwise differences $\Delta\Delta G_{j,k} = \Delta G_j - \Delta G_k$:

*Medicine Design, Worldwide Research & Development, Pfizer Inc. 1 Portland St, Cambridge MA 02139, United States

†Early Clinical Development, Worldwide Research & Development, Pfizer Inc. 445 Eastern Point Rd, Groton CT 06340, United States

‡XtalPi Inc. One Broadway, Cambridge, MA 02142, United States

§Department of Chemistry, University of California, Irvine. 3134B Natural Sciences I, Irvine, CA 92697, United States

$$\mathbf{A}_{\Delta\Delta G} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & . & . & . & . & . & 0 \\ 0 & . & . & . & . & . & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}_{(m^{FEP},N)} .$$

The matrix **A** defines the **design topology** of pairwise-based calculations.

Computer-Generated Designs

To generate a randomly selected design, m_{FEP} unique pairs were randomly selected from ${}_NC_2 = N(N-1)/2$ total pairs to form matrix A. The rank of matrix A must be equal to $N - m_{exp}$ in order to solve specific $\hat{\Delta G}$ through equation eq (2).

Both D- and A-optimality use function of eq (3) to obtain the appropriate design that is defined by the scaled moment matrix, M, expressed as

$$M = \frac{I}{N\sigma^2} = \frac{(\mathbf{A}' W^{-1} \mathbf{A})^{-1}}{N}$$

where I is the inverse of *Information Matrix* as referred in the main text. The Fedorov-exchange algorithm finds that set of design points that minimizes the determinant of M^{-1} for D-optimality, and minimizes the trace of M^{-1} for A-optimality.

Simulation Study

In our simulation study, the matrix **A** represents a **Topological Design**. Each simulation, from step 1 to 6 represents a new independent **Experiment**. Figure S1 exemplifies the distinction between different designs and experiments. We believe that this is a better approach to study the role of topological design in prediction accuracy because it separates the variability from target and ligand selection, and, therefore, can be applied to a broader range of pairwise-based calculations.

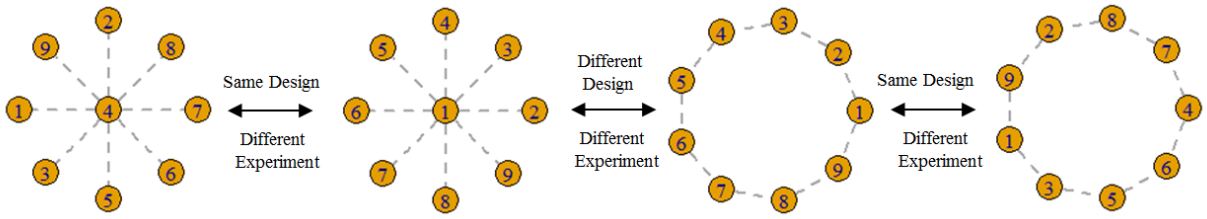

Figure S1: Visualization of different designs and experiments. Each node represents a specific ligand labeled by number. Dashed lines represent the presence of FEP calculated $\Delta\Delta G^{FEP}$ values between pairs of ligands.

Example of Simulation Based on One Specific Design

Here we show one example of a simulation based on the A-optimal design of 30 pairs and 20 ligands presented in Figure S2. For this example, one ligand has an experimentally known ΔG^{True} value and is used as the reference ligand. A total of 30 pairs were selected based on A-optimality. The $\Delta\Delta G^{FEP}$ is simulated 5000 times with normally distributed errors using function rnorm(mean = $\Delta\Delta G^{True}$, sd=$RMSE_{\Delta\Delta G^{FEP}}$) in R, where $RMSE_{\Delta\Delta G^{FEP}}$ is set to 1.0 (kcal/mol). Figure S2 shows the distribution of the observed MSE of $\Delta\Delta G^{FEP}$ vs. $\Delta\Delta G^{True}$, the observed MSE of the transformed $\hat{\Delta G}$ estimates vs. ΔG^{True}, as well as the observed Spearman correlation ρ between the back transformed $\hat{\Delta G}$ vs ΔG^{True}.

For this design example, the MSE between $\Delta\Delta G^{FEP}$ and $\Delta\Delta G^{True}$, as expected, is almost normally distributed with median value at 1.0 kcal/mol and 95% quantiles of the MSE distribution at (0.56-1.54), consistent with the value used for simulation. The back transformed $\hat{\Delta G}$ has median MSE of 0.80 kcal/mol and 95% quantiles of (0.28-3.39). The observed median Spearman's correlation ρ between $\hat{\Delta G}$ estimates and ΔG^{True} is 0.79 with 95% quantiles of (0.52-0.93). As described in methods section, each simulation represents a new and independent **experiment**. Because the MSE of $\Delta\Delta G^{FEP}$ vs $\Delta\Delta G^{True}$ is an intrinsic variable in simulation, it is solely dependent on the intrinsic accuracy of Free Energy Perturbation calculation itself and independent of pair and reference ligand selection. However, the distribution of $MSE_{\hat{\Delta G}}$ is a function of both the experimental design and intrinsic accuracy of FEP calculations.

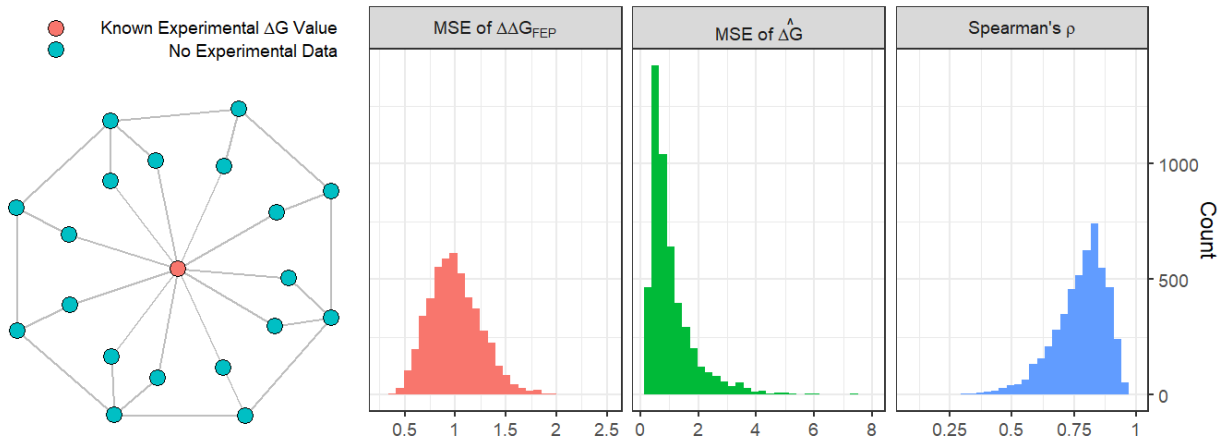

Figure S2: Example simulation of an A-optimal design with 30 perturbation pairs of 20 ligands (1 reference and 19 unknown). Left, visualization of the A-optimal design. The reference ligand is colored green. Right, distribution of $MSE_{\Delta\Delta G^{FEP}}$, $MSE_{\hat{\Delta G}}$, and Spearman's correlation ρ between $\hat{\Delta G}$ estimate and ΔG^{True}.

Summary of Simulations with fixed ΔG^{True} values

In order to represent the same ligand set scenario, the fixed ΔG^{True} values were also studied in simulation. In this case, ΔG^{True}s from the motivating example (Table 1) were used with ligand 1 (ΔG^{True} = -8.09) as the reference in all the simulations. The corresponding $\Delta\Delta G^{FEP}$ was generated the same way as described in the methods section.

It is clear that there is no difference in Mean Squared Error (MSE)-based metrics compared to the values obtained from randomly generated, normally distributed ΔG^{True}: both the averaged and theoretically calculated ones. This indicates, again, that the input $MSE_{\Delta\Delta G^{FEP}}$ and the design are the major contributing factors to the MSEs. The Spearman correlation ρ is overall better mostly because the ΔG^{True} values from the motivating example has larger scale ($\Delta G^{True}_{Max} - \Delta G^{True}_{Min}$ = 5.89 kcal/mol) than the average scale of the randomly generated ΔG^{True} values, which is around 3.79 based on 5000 fold sampling using function rnorm(n=20, mean=0, sd=1.0). It is important to note that, although the correlation appears better in this simulation, the accuracy of identifying the best ligand still remains in the same range, or event worse across all the designs.

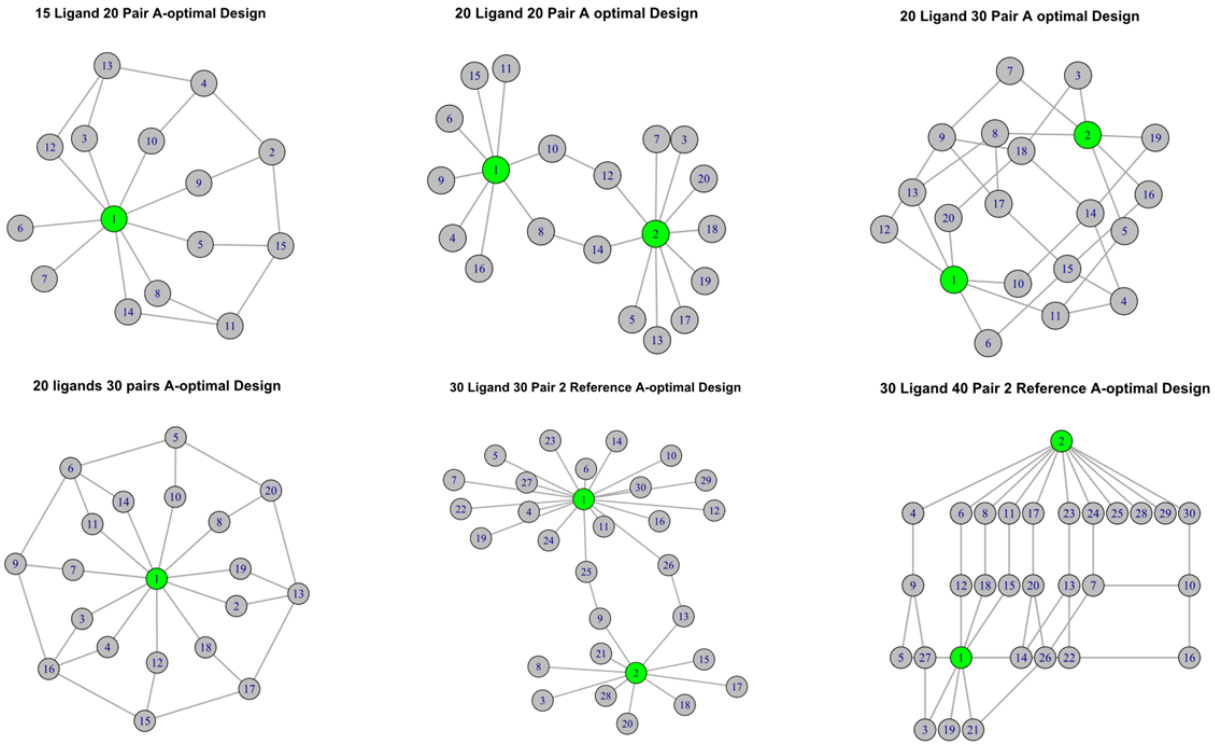

Figure S3: Example of A-optimal designs with different number of ligands, references and pairs. Each solid line represents an available FEP calculation $\Delta\Delta G^{FEP}$ value between the connecting ligands. Green nodes represent the reference ligands with known experimental values ΔG^{Exp}

Pairs	Design	$MSE_{\hat{\Delta G}}$ [a]	$MSE_{\Delta\hat{\Delta G}}$ [b]	$MSE^{All}_{\Delta\hat{\Delta G}}$ [c]	Median ρ [d]	Accuracy[e]
20	Random*	4.00	0.95	3.80	0.81 (0.68-0.89)	27.0%
	D-optimal	3.83	0.95	3.50	0.82 (0.69-0.91)	28.3%
	A-optimal	1.42	0.95	1.83	0.88 (0.82-0.92)	30.8%
30	Random*	1.66	0.63	1.33	0.91 (0.85-0.95)	33.0%
	D-optimal	1.36	0.63	0.91	0.93 (0.89-0.96)	36.2%
	A-optimal	1.08	0.63	1.10	0.92 (0.88-0.95)	37.1%
50	Random*	0.98	0.38	0.54	0.95 (0.93-0.97)	42.6%
	D-optimal	0.92	0.38	0.44	0.96 (0.94-0.98)	45.8%
	A-optimal	0.79	0.38	0.50	0.96 (0.94-0.97)	43.7%

* The averages across 5,000 different randomly selected designs.

[a] $MSE_{\hat{\Delta G}}$ was theoretically derived from equation (3) for D- and A-optimal design.

[b] $MSE_{\Delta\hat{\Delta G}}$ was theoretically derived from equation (3) for D- and A-optimal design for the corresponding FEP pairs.

[c] $MSE^{All}_{\Delta\hat{\Delta G}}$ was theoretically derived from equation (3) for D- and A-optimal of all ${}_NC_2 = 190$ pairs.

[d] Spearman ρ was calculated between ΔG^{True} and estimated $\hat{\Delta G}$ value of 20 ligands. 15% - 85% quantiles are in parenthesis.

[e] Probability of correctly identifying the best ligand.

Table S1: Analytically derived and simulated metrics for different designs with fixed ΔG^{True} from the motivating example in introduction section. The simulated $\Delta\Delta G^{FEP}$ values were generated in R using function rnorm(mean = $\Delta\Delta G^{True}$, sd = sqrt($MSE_{\Delta\Delta G^{FEP}}$)), where $N = 20$ ligands and $MSE_{\Delta\Delta G^{FEP}} = 1.0$ kcal/mol.

Comparison of Mean Squared Error (MSE)

In Figure S4, we compared the mean squared error (MSE) of the FEP calculated $\Delta\Delta G^{FEP}$ values (orange boxplots) with the MSE of the pairwise differences of $\hat{\Delta G}$ estimates, also

referred to as $\Delta\hat{\Delta}G$s, (blue boxplots) that had corresponding $\Delta\Delta G^{FEP}$ values. Note that the MSEs for each of these quantities may also be computed analytically. We have included the analytical estimates on the plot with a yellow triangle. The boxplots corresponding to the simulation study give both confirmation of the theoretical MSE results and an idea of the uncertainty present in the MSE statistics for a data set with 20 ligands and variability similar to the simulation parameters. Observe that the back-transformed $\Delta\hat{\Delta}G$ estimates are uniformly more accurate than the original, FEP calculated $\Delta\Delta G^{FEP}$ values, as one would expect when incorporating additional information into the estimates. Also note that there does not appear to be any difference in average MSE of $\Delta\hat{\Delta}G$ with respect to the design. This will change when we consider the set of all possible $\Delta\hat{\Delta}G$ estimates rather than only pairwise differences that have corresponding FEP calculated $\Delta\Delta G^{FEP}$ values, as we will discuss in the next figure S5. Lastly, as expected, when more pairs are included in the model, the back transformed estimates become more precise (theoretical MSEs of 0.95 with 20 pairs, 0.63 with 30 pairs, and 0.38 with 50 pairs).

In Figure S5, we compared MSEs of the set of all possible pairwise differences of $\hat{\Delta G}$ estimates ($\Delta\hat{\Delta G}_{i,j} = \hat{\Delta G}_i - \hat{\Delta G}_j$) across the different designs. Note that one cannot estimate all possible pairs using only $\Delta\Delta G^{FEP}$ values without back transforming to $\hat{\Delta G}$ estimates. At least one reference ligand is needed for back transformation. Observe that the D-optimal design performs quite poorly, using this metric, when only 20 pairs are included (theoretical MSE of 3.50 vs 1.83 for the A- and D-optimal designs). The D-optimal design starts to perform better as the number of pairs increase, however, eventually out performing the A-optimal design. Additionally, as one would expect with this criteria, the D-optimal design performs best for all 3 design sizes. Alternatively, the randomly generated designs perform the worst across all 3 design sizes, indicating the benefits of choosing a design according to an optimality criterion.

In Figure S6 and S7, we computed the MSEs and Rank Correlation distribution of the back-transformed $\hat{\Delta G}$ estimates themselves, rather than their pairwise differences. The A-optimality criterion was designed to perform best under this metric, which it does under all 3 design sizes we considered. Additionally, as we saw in the previous graphic, the D-optimal

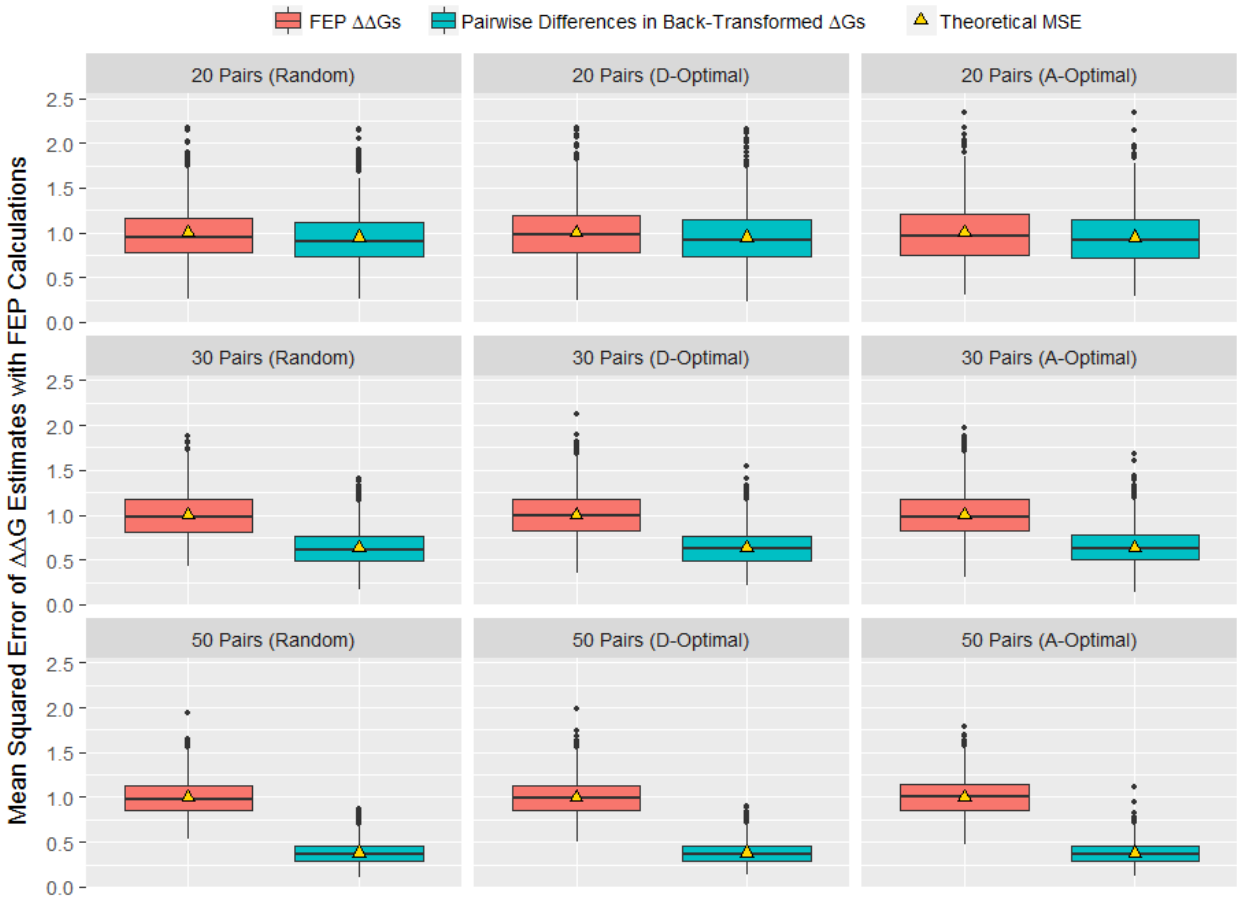

Figure S4: Mean squared error of $\Delta\Delta G$ estimates. The orange boxplots represent simulated MSE estimates for FEP derived $\Delta\Delta G^{FEP}$ estimates while the blue boxplots represent the simulated MSE estimates for the pairwise differences of $\hat{\Delta G}$ estimates with corresponding FEP values. The theoretical MSEs are indicated by yellow triangles.

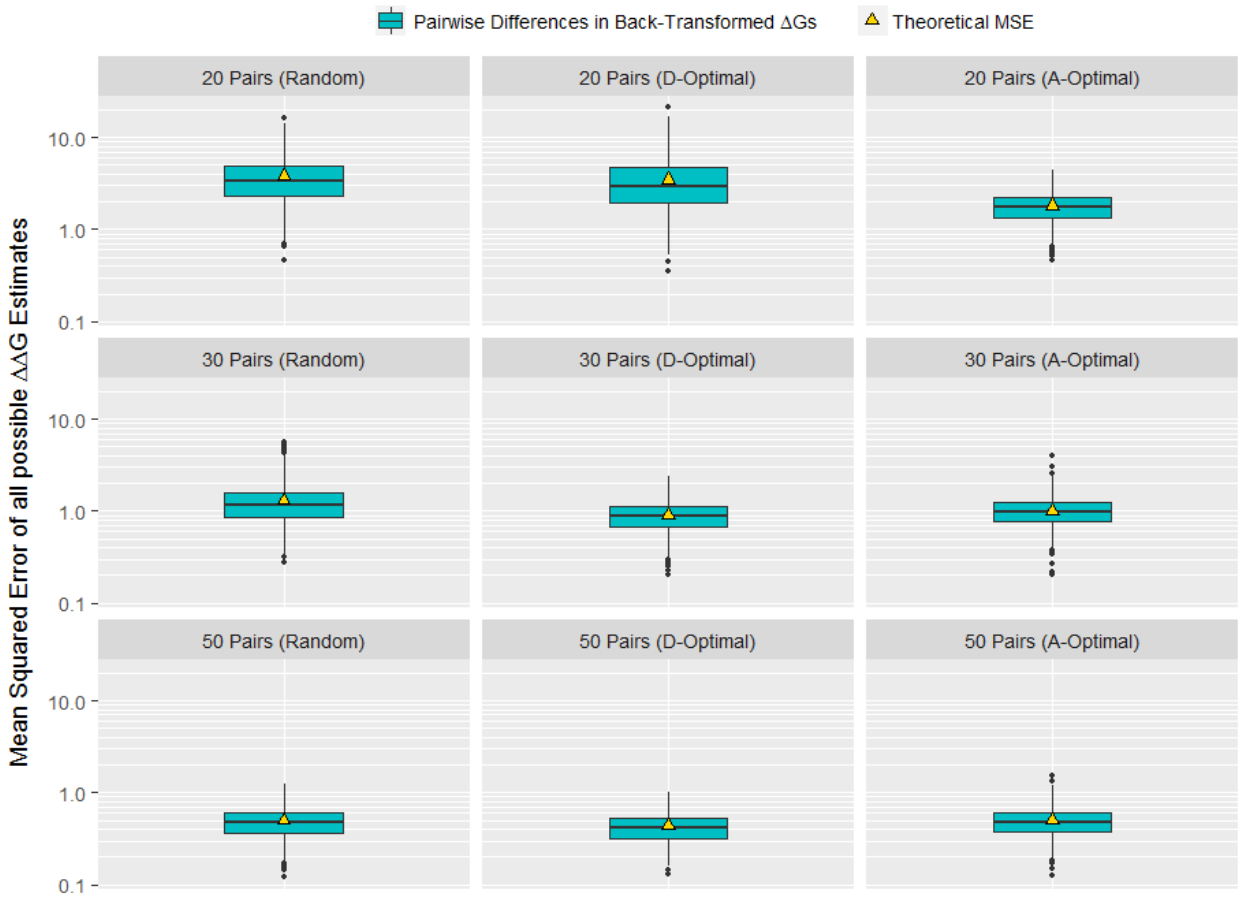

Figure S5: Mean squared error of $\Delta\Delta G$ estimates for all possible pairs. The blue boxplots represent the simulated MSE estimates for all possible pairwise differences of $\hat{\Delta G}$ estimates. The theoretical MSEs are indicated by yellow triangles.

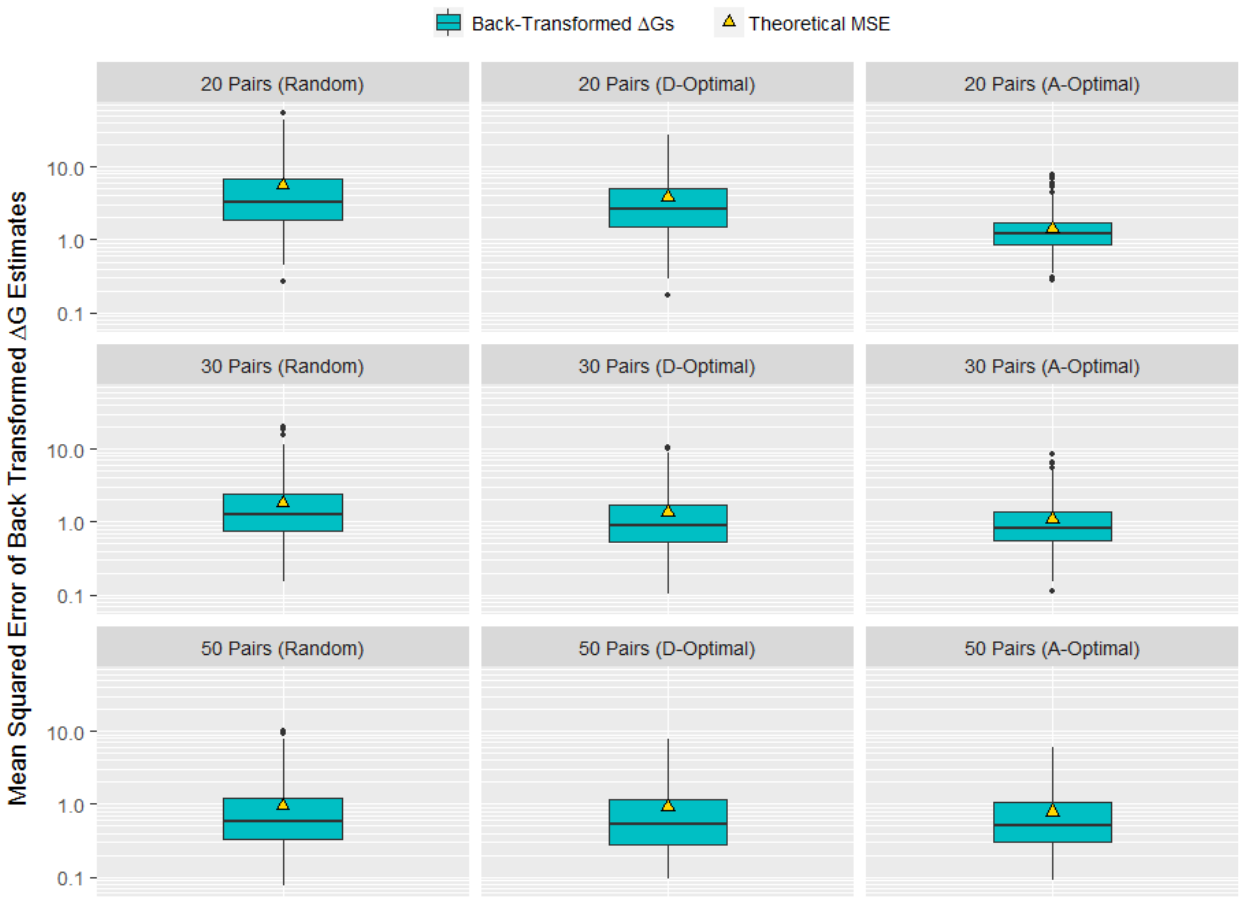

Figure S6: Mean squared error of back transformed ΔG estimates. The blue boxplots represent the simulated MSE estimates for the back transformed ΔG estimates. The theoretical MSEs are indicated by yellow triangles.

design performs quite poorly here, and is the worst of the 3 designs selected via an optimality criterion across all 3 design sizes, only out-performing the randomly generated designs. Again, this discrepancy was largest when only 20 pairs were included in the design, indicating that structuring your pairs in a ring is an extremely poor choice. Lastly, the deficiencies of the randomly selected design once again emphasize the improvements in accuracy that may be achieved by constructing the study designs intelligently.

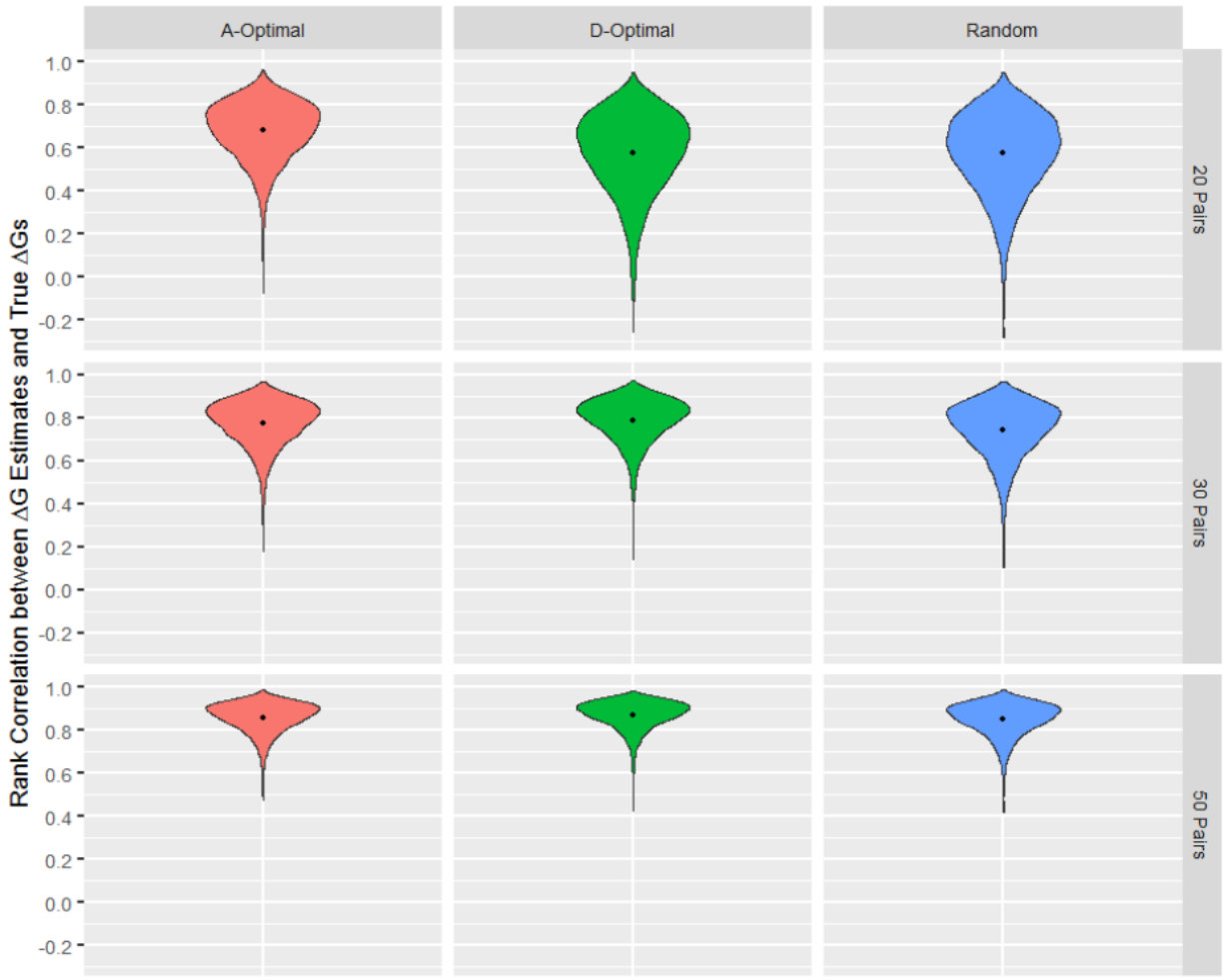

Figure S7: Distribution of Spearman's Rank Correlation for each design. The center points represent the mean values while the distribution represents all values generated across the 5,000 simulations.

Optimal Designs with Structure-Similarity Based Weighting

The AtomPair fingerprint based tanimoto score is calculated for all the pairs among 16 CDK2 ligands. The weighing factor is the normalized tanimoto score defined as:

$$w_{i,j} = TanimotoScore^2_{i,j}/(Max(TanimotoScore^2_{i,j}) - Min(TanimotoScore^2_{i,j}))$$

The normalization is to further differentiate the highly similar ligands characterized by the fingerprint. Another example of weighted optimal design without scaling the similarity scores is demonstrated here as well. Figure S8 shows the examples of weighted A- and D-optimal designs using ECFP_2[1,2] fingerprint Tanimoto score as the weighting factor. It is clear that, compared to unweighted optimal or the literature design, the similarity-weighted design includes more pairs with high structural similarity. Ligand CAT-4o, was selected as reference since, according to literature design, that specific ligand is critical in connecting two sub perturbation graphs.

The relative design efficiency of different optimal designs is listed in Table S2. It is worthwhile noting that, for this example, the relative efficiency for the weighted designs is slightly higher than the corresponding unweighted. This is because the weighted designs are biased toward the pairs with higher structural similarity, which are believed to have smaller errors. Therefore, the weighted optimality criteria is better than the unweighted ones, e.g. $det(I^{-1}_{weighted}) \leq det(I^{-1}_{unweighted})$.

Free Energy Perturbation Calculation

CDK2 protein and ligand structures are from reference 3.[3] Amber18 GPU-TI code[4] was used to conduct the alchemical free energy calculations. Similar to previously reported protocol,[5] dual-topology was used to generate TI topology. A softcore potential was applied to appearing and disappearing atoms in the two end states. Gaff[6] version 2 in Amber18 was used as small molecule force filed with Amber FF14SB[7] as protein force field. For TI simulation, the systems were solvated in a water box with buffer width of 10 Å for both complex and solvent simulations. The systems were minimized and equilibrated using the

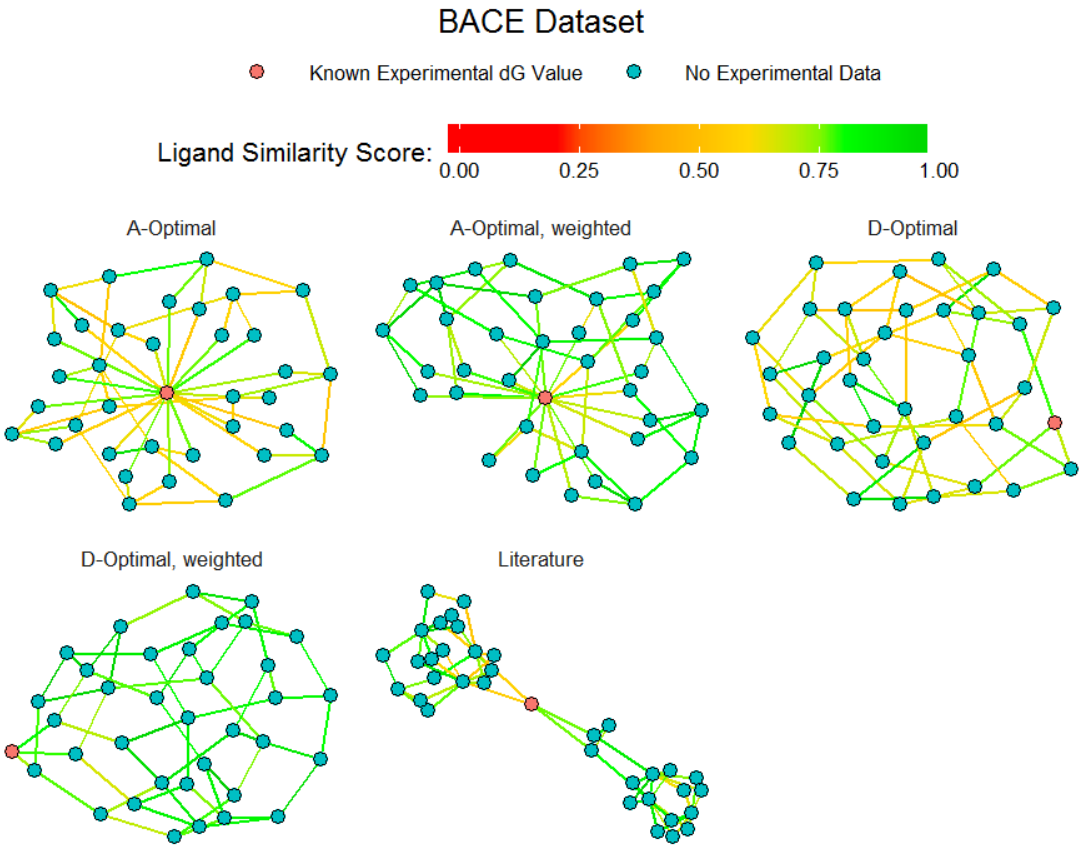

Figure S8: Example designs of the BACE ligand set from reference[3]. One ligand (CAT-4o) is selected as reference and colored in red for illustration purpose. Each edge is colored with fingerprint Tanimoto score of the connecting pair. All the designs have the same reference ligand and the same number of pairs. Specific reference and pairs are listed in the SI table.

A-Optimal	Weighted A-Optimal	D-Optimal	Weighted D-optimal
1.57	1.73	1.18	1.31

[a] relative efficiency of the optimal design is calculated as the ratio of the corresponding criteria of literature design to its A-Optimality or D-Optimality criteria: $tr(I^{-1}_{literature})/tr(I^{-1}_{A-optimal})$ and $det(I^{-1}_{literature})/det(I^{-1}_{D-optimal})$.

Table S2: Relative efficiency of optimal designs compared to literature design of BACE compounds.[3]

default setting: the whole system with the solute molecules was restrained to their initial positions and was first minimized using steepest descent method and then simulated at 298 K using an NVT ensemble with the restraint retained for 20 ps. After that the system was simulated at room temperature using the NPT ensemble without any restraint for 100 ps followed by the production simulation for 2 ns for both complex and solvent systems. A total number of 10 stages and 22 λ windows for each stage were used to achieve good convergence for this CDK2 system. The total production simulation time is 2 ns for both the complex and the solvent simulations. Only data in the last 1ns production NPT run was used in the analysis. We used alchemlyb[8] to analyze the result.

References

1. D. Rogers and M. Hahn, Journal of Chemical Information and Modeling **50**, 742 (2010), ISSN 1549-9596, URL https://doi.org/10.1021/ci100050t.

2. *Biovia pipeline pilot 9.0* (2013).

3. L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, et al., J Am Chem Soc **137**, 2695 (2015), ISSN 0002-7863, URL http://pubs.acs.org/doi/pdfplus/10.1021/ja512751q.

4. T.-S. Lee, Y. Hu, B. Sherborne, Z. Guo, and D. M. York, Journal of Chemical Theory and Computation **13**, 3077 (2017), ISSN 1549-9618, URL https://doi.org/10.1021/acs.jctc.7b00102.

5. H. H. Loeffler, J. Michel, and C. Woods, Journal of Chemical Information and Modeling **55**, 2485 (2015), ISSN 1549-9596, URL https://doi.org/10.1021/acs.jcim.5b00368.

6. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J Comput Chem **25**, 1157 (2004), ISSN 0192-8651 (Print) 0192-8651.

7. J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, Journal of Chemical Theory and Computation **11**, 3696 (2015), ISSN 1549-9618, URL https://doi.org/10.1021/acs.jctc.5b00255.

8. D. Dotson and I. Kenney, *alchemistry/alchemlyb: Release 0.1.0.a1* (2017), URL https://doi.org/10.5281/zenodo.293736.