Is structure based drug design ready for selectivity optimization?

Steven K. Albanese[1,2,†], John D. Chodera[2], Andrea Volkamer[3], Simon Keng[4], Robert Abel[4], Lingle Wang[4*]

[1]Louis V. Gerstner, Jr. Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer Center, New York, NY 10065; [2]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065; [3]Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin; [4]Schrödinger, New York, NY 10036

***For correspondence:** lingle.wang@schrodinger.com (LW)

Present address: [†]Schrödinger, New York, NY 10036

Abstract

Alchemical free energy calculations are now widely used to drive or maintain potency in small molecule lead optimization with a roughly 1 kcal/mol accuracy. Despite this, the potential to use free energy calculations to drive optimization of compound *selectivity* among two similar targets has been relatively unexplored in published studies. In the most optimistic scenario, the similarity of binding sites might lead to a fortuitous cancellation of errors and allow selectivity to be predicted more accurately than affinity. Here, we assess the accuracy with which selectivity can be predicted in the context of small molecule kinase inhibitors, considering the very similar binding sites of human kinases CDK2 and CDK9, as well as another series of ligands attempting to achieve selectivity between the more distantly related kinases CDK2 and ERK2. Using a Bayesian analysis approach, we separate systematic from statistical error and quantify the correlation in systematic errors between selectivity targets. We find that, in the CDK2/CDK9 case, a high correlation in systematic errors suggests free energy calculations can have significant impact in aiding chemists in achieving selectivity, while in more distantly related kinases (CDK2/ERK2), the correlation in systematic error suggests fortuitous cancellation may even occur between systems that are not as closely related. In both cases, the correlation in systematic error suggests that longer simulations are beneficial to properly balance statistical error with systematic error to take full advantage of the increase in apparent free energy calculation accuracy in selectivity prediction.

Free energy methods have proven useful in aiding structure-based drug design by driving the optimization or maintenance of potency in lead optimization. Alchemical free energy calculations allow for the prediction of ligand binding free energies, including all enthalpic and entropic contributions [1]. Advances in atomistic molecular mechanics simulations and free energy methodologies [2–5] have allowed free energy methods to reach a level of accuracy sufficient for predicting ligand potencies [6]. These methods have been applied prospectively to develop inhibitors for Tyk2 [7], Syk [8], BACE1 [9], GPCRs [10], and HIV protease [11]. A recent large-scale review of the use of FEP+ [12] to predict potency for 92 different projects and 3 021 compounds determined that predicted binding free energies had a median root mean squared error (RMSE) of 1.0 kcal/mol [13].

Selectivity is an important consideration in drug design

In addition to potency, selectivity is an important property to consider in drug development, either in the pursuit of an inhibitor that is maximally selective [14, 15] or possesses a desired polypharmacology [16–

20]. Controlling selectivity can be useful not only in avoiding off-target toxicity (arising from inhibition of unintended targets) [21, 22], but also in avoiding on-target toxicity (arising from inhibition of the intended target) by selectively targeting disease mutations [23]. In either paradigm, considering the selectivity of a compound is complicated by the biology of the target. For example, kinases exist as nodes in complex signaling networks [24, 25] with feedback inhibition and cross-talk between pathways. Careful consideration of which off-targets are being inhibited can avoid off-target toxicity due to alleviating feedback inhibition and inadvertently reactivating the targeted pathway [24, 25] or the upregulation of a secondary pathway by alleviation of cross-talk inhibition [26, 27]. Off-target toxicity can also be caused by inhibiting unrelated targets, such as gefitinib, an EGFR inhibitor, inhibiting CYP2D6 [21] and causing hepatotoxicity in lung cancer patients. In a cancer setting, on-target toxicity can be avoided by considering the selectivity for the oncogenic mutant form of the kinase over the wild type form of the kinase [28–30], exemplified by a number of first generation EGFR inhibitors. Selective binding to multiple kinases can also lead to beneficial effects: Imatinib, initially developed to target BCR-Abl fusion proteins, is also approved for treating gastrointestinal stromal tumors (GIST) [31] due to its activity against receptor tyrosine kinase KIT.

The use of physical modeling to predict selectivity is relatively unexplored

While engineering compound selectivity is important for drug discovery, the utility of free energy methods for predicting this selectivity with the aim of reducing the number of compounds that must be synthesized to achieve a desired selectivity profile has been relatively unexplored in published studies. If there is fortuitous cancellation of systematic errors for closely related systems, free energy methods may be much more accurate than expected given the errors made in predicting the potency for each individual target. Such systematic errors might arise from force field parameters uncertainty, force field parameters assignment, protein or ligand protonation state assignment, or even from systematic errors arising in the target experimental data, where for example poor solubility of a particular compound might lead to a spuriously poor reported binding affinity for that compound, among other effects.

Molecular dynamics and free energy calculations have been used extensively to investigate the biophysical origins of the selectivity of imatinib for Abl kinase over Src [32, 33] and within a family of non-receptor tyrosine kinases [34]. This work focused on understanding the role reorganization energy plays in the exquisite selectivity of imatinib for Abl over the highly related Src despite high similarity between the cocrystallized binding mode and kinase conformations, and touches on neither the evaluation of the accuracy of these methods nor their application to drug discovery on congeneric series of ligands. Previous work predicting the selectivity of three bromodomain inhibitors across the bromodomain family achieved promising accuracy for single target potency of roughly 1 kcal/mol, but does not explicitly evaluate any selectivity metrics [35] or quantify the correlation in the errors made in predicting affinities for each bromodomain. Previous work using FEP+ to predict selectivity between pairs of phosphodiesterases (PDEs) showed promising performance but did not evaluate correlation in the error made in predicting affinities for each PDE [36]

Kinases are an important and particularly challenging model system for selectivity predictions

Kinases are a useful model system to work with for assessing the utility of free energy calculations to predict inhibitor selectivity in a drug discovery context. With the approval of imatinib for the treatment of chronic myelogenous leukemia in 2001, targeted small molecule kinase inhibitors (SMKIs) have become a major class of therapeutics in treating cancer and other diseases. Currently, there are 52 FDA-approved SMKIs [37], and it is estimated that kinase targeted therapies account for as much as 50% of current drug development [38], with many more compounds currently in clinical trials. While there have been a number of successes, the current stable of FDA-approved kinase inhibitors targets only a small number of kinases implicated in disease, and the design of new selective kinase inhibitors remains a significant challenge.

Achieving selective inhibition of kinases is quite challenging, as there are more than 518 protein kinases [39, 40] sharing a highly conserved ATP binding site that is targeted by the majority of SMKIs [41]. While kinase inhibitors have been designed to target kinase-specific sub-pockets and binding modes to achieve selectivity [42–47], previous work has shown that both Type I (binding to the active, DFG-in conformation) and Type II (binding to the inactive, DFG-out conformation) inhibitors are capable of achieving a

range of selectivities [48, 49], often exhibiting significant binding to a number of other targets in addition to their primary target. Even FDA-approved inhibitors—often the result of extensive drug development programs—bind to a large number of off-target kinases [50]. Kinases are also targets of interest for developing polypharmacological compounds, or inhibitors that are specifically designed to inhibit multiple kinase targets. Resistance to MEK inhibitors in KRAS-mutant lung and colon cancer has been shown to be driven by ErbB3 upregulation [51], providing a rationale for dual MEK/ERBB family inhibitors. Similarly, combined MEK and VEGFR1 inhibition has been proposed as a combinatorial approach to treat KRAS-mutant lung cancer [52]. Developing inhibitors with a desired polypharmacology means navigating more complex selectivity profiles, presenting a problem where physical modeling has the potential to dramatically speedup drug discovery.

The correlation coefficient measures how useful predictions are in achieving selectivity

Since the prediction of selectivity depends on predicting affinities to two or more targets (or relative affinities between pairs of related molecules), a spectrum of possibilities exists for how accurately selectivity can be predicted even when the error in predicting individual target affinities is fixed. In well-behaved kinase systems, for example, free energy calculation potency predictions have achieved root-mean-square of less than 1.0 kcal/mol [7, 12]. This residual error likely arises from a variety of contributions. Systematic contributions to the residual error may include forcefield parameterization deficiencies, protein and ligand protonation assignment errors, and discrepancies between the crystallographic construct protein and the assay construct protein. Likewise, unbiased contributions to the observed residual error likely arises from incompletely converged sampling. Lastly, it should not be forgotten that the target experimental value will have its own systematic and random errors.

In the best-case scenario, correlation in the systematic errors for predicting the interactions of a given ligand with two related protein targets might exactly cancel out, allowing selectivity to be predicted much more accurately than potency. On the other hand, if the uncorrelated random error dominates the residual error between two protein targets, predictions of selectivity will be *less accurate* than potency predictions. Real-world systems are likely to fall somewhere between these two extremes, and quantifying the *degree* to which error in multiple protein targets is correlated, its implications for the use of free energy calculations for prioritizing synthesis in the pursuit of selectivity, the ramifications for optimal calculation protocols, and rough guidelines governing which systems we might expect good selectivity prediction is the primary focus of this work.

In particular, in this work, we investigate the magnitude of the correlation (ρ) in error for the predicted binding free energy differences between related compounds ($\Delta\Delta G_{ij}$) for two different targets, assessing the utility of alchemical free energy calculations for the prediction of selectivity. We employ state of the art relative free energy calculations [12, 13] to predict the selectivities of two different congeneric ligand series [53, 54], and construct simple numerical models that allow us to quantify the potential utility in selectivity optimization expected for different combinations of per target systematic errors and correlation coefficients. To make a realistic assessment of our confidence in this correlation coefficient derived from a limited number of experimental measurements, we develop a new Bayesian approach to quantify the uncertainty in the correlation coefficient in the predicted change in selectivity on ligand modification, incorporating all sources of uncertainty and correlation in the computation to separate statistical from systematic error. We find that in the closely related systems of CDK2 and CDK9, a high correlation of systematic errors suggests that free energy methods can have a significant impact on speeding up selectivity optimization. Even in the more distantly related case (CDK2/ERK2), correlation in the systematic errors allows free energy calculations to speedup selectivity optimization, suggesting that these methodologies can impact drug discovery even when comparing systems that are less closely related. We also present a model of the impact of per target statistical error at different levels of systematic error correlation, suggesting that it is worthwhile to expend more effort sampling in systems with high correlation.

Results

Alchemical free energy methods can be used to predict compound selectivity

While the potency of a ligand i for a single target is often quantified as a free energy of binding ($\Delta G_{i,\text{target}}$), there are a number of different metrics for quantifying compound selectivity [55, 56]. Here, we consider the selectivity S_i between one target and another (an *antitarget*) as the difference in free energy of binding for a given ligand i between the two,

$$S_i \equiv \Delta G_{i,\text{target 2}} - \Delta G_{i,\text{target 1}} \tag{1}$$

While in the optimization of potency we are concerned with $\Delta\Delta G_{ij} \equiv \Delta G_j - \Delta G_i$, the relative free energy of binding of ligands i and j to a single target, in the optimization of selectivity, we are concerned with $\Delta S_{ij} \equiv S_j - S_i$, which reflects the change in selectivity between ligand i and a related ligand j,

$$\begin{aligned} \Delta S_{ij} &\equiv S_j - S_i \\ &= (\Delta G_{\text{j, target 2}} - \Delta G_{\text{j, target 1}}) - (\Delta G_{\text{i, target 2}} - \Delta G_{\text{i, target 1}}) \\ &= \Delta\Delta G_{\text{ij, target 2}} - \Delta\Delta G_{\text{ij, target 1}} \end{aligned} \tag{2}$$

To predict the change in selectivity, ΔS_{ij}, between two related compounds, we developed a protocol that uses a relative free energy calculation (FEP+) [12] to construct a map of alchemical perturbations between ligands in a congeneric series, as described in detail in the **Methods**. The calculation is repeated for each target of interest, with identical perturbations (edges) between each ligand (nodes). Each edge represents a relative alchemical free energy calculation that quantifies the $\Delta\Delta G$ between the ligands (nodes) for a single target. From these calculations, we can then compute the change in selectivity between the two targets of interest, ΔS_{ij}, achieved by transforming ligand i into ligand j.

Previous work has demonstrated that FEP+ can achieve an accuracy (σ_{target}) of roughly 1 kcal/mol in single-target potency prediction, which reflects a combination of systematic systematic error and random statistical error [12]. However, it is possible that the systematic error for a given perturbation between ligands i and j ($\sigma_{\text{sys, ij, target}}$) in two different systems may fortuitously cancel when computing ΔS_{ij}, resulting in the systematic contribution to the selectivity error ($\sigma_{\text{selectivity}}$) being significantly lower than its contribution to single-target potency error (σ_{target}). This systematic error may cancel between the two systems for a variety of reasons. For example, a ligand force field parameter assignment error might lead to an spuriously large solvation free energy for a particular compound, which will cancel in the selectivity analysis. Likewise, a sparingly soluble compound might have a similar experimental measurement error for the on-target protein as the off-target protein. Similar cancellation of systematic errors might be observed in ligand and/or protein protonation state assignment error, or systematic differences existing between the protein constructs used for crystallographic studies and biochemical or biophysical assays.

If we presume that the systematic errors for both targets are distributed according to a bivariate normal distribution with correlation coefficient ρ quantifying the *degree* of correlation (with $\rho = 0$ denoting no correlation and $\rho = 1$ denoting perfect correlation), and that the statistical errors for both targets ($\sigma_{\text{stat,ij,target}}$) are completely independent, we can model the error in predicting the ΔS_{ij} as $\sigma_{\text{selectivity}}$,

$$\sigma_{\text{selectivity}} \equiv \sqrt{\sigma^2_{\text{sys,ij,1}} + \sigma^2_{\text{sys,ij,2}} - 2\rho\,\sigma_{\text{sys,ij,1}}\,\sigma_{\text{sys,ij,2}} + \sigma^2_{\text{stat,ij,1}} + \sigma^2_{\text{stat,ij,2}}} \tag{3}$$

$\sigma_{\text{selectivity}}$ can be split into two components: systematic error and statistical error. As more effort is spent sampling, the per-target statistical error for a given transformation from ligand i to ligand j ($\sigma_{\text{stat,ij, target}}$) will decrease, eventually becoming zero in the regime of infinite sampling. As we shall see below, the quantitative value of the correlation coefficient ρ for the systematic error component has important ramifications for the accuracy with which selectivity can be predicted.

Correlation in systematic errors can significantly enhance accuracy of selectivity predictions

To demonstrate the potential impact the correlation coefficient ρ has on predicting selectivity using alchemical free energy techniques, we created a simple numerical model following Equation 3 which takes into account each of the per-target systematic errors ($\sigma_{\text{sys,ij,1}}, \sigma_{\text{sys,ij,2}}$) expected from the methodology as well as the correlation in those errors, while assuming infinite effort is spent sampling to reduce the statistical error component (σ_{stat}) to zero. As seen in Figure 1A, if the per target systematic errors are the same magnitude

($\sigma_{sys,ij,1} = \sigma_{sys,ij,2}$), $\sigma_{selectivity}$ approaches 0 as the correlation coefficient ρ approaches 1, even though the single-target potency systematic error is nonzero. If the error for the free energy method is not the same magnitude ($\sigma_{sys,ij,1} \neq \sigma_{sys,ij,2}$), $\sigma_{selectivity}$ gets smaller but approaches a non-zero value as ρ approaches 1.

To quantify the expected reduction in number of compounds that must be synthesized to achieve a desired selectivity threshold (hereafter referred to as the *speedup* in selectivity optimization), we modeled the change in selectivity with respect to a reference compound for a number of compounds a medicinal chemist might suggest as a normal distribution centered around 0 with a standard deviation of 1 kcal/mol (Figure 1B, black curve), reflecting the notion that most proposed modifications would not drive large changes in selectivity. This assumption—that a synthetic chemist's proposal distribution can be modeled as a normal distribution—is based on data-driven estimates from an Abbott Laboratories analysis of potency changes [57]

Further suppose that each compound is evaluated computationally with a free energy methodology that has a per-target systematic error ($\sigma_{sys,\ ij,target}$) of 1 kcal/mol, where we presume sufficient computational effort has been expended to make statistical error negligible. All compounds predicted to have a 1.4 kcal/mol or greater improvement in selectivity (10× in ratio of affinities, or 1 $\log_{10}$ unit) are synthesized and experimentally tested (Figure 1B, colored curves), using an experimental technique with perfect measurement accuracy. The fold-change in the proportion of compounds that are made that have a true 1.4 kcal/mol improvement in selectivity compared to the original distribution can be calculated as a surrogate for the expected speedup. For this 1.4 kcal/mol selectivity improvement threshold, a correlation coefficient ρ = 0.5 gives an expected speedup of 4.1×, which can be interpreted as needing to make 4.1x fewer compounds to achieve a tenfold improvement in selectivity. This process can be extended for the even more difficult proposition of achieving a hundredfold improvement in selectivity (Figure 1C), where 200–300× speedups can be expected, depending on the single-target systematic error ($\sigma_{sys,ij,target}$) for the free energy methodology.

These estimates represent an ideal scenario, where the number of compounds scored and synthesized is unlimited. In a more realistic discovery project, the number of compounds scored is limited by computational resources, and the number of compounds synthesized is limited by chemistry resources. In this case, the observed speedup will depend not only on the correlation coefficient ρ and per-target systematic error ($\sigma_{sys,ij,target}$), but also the number of compounds scored and the synthesis rule, defined as the selectivity threshold a compound must be predicted to reach before being selected for synthesis. To model this process, suppose a given number of compounds (Figure 1D, x-axis of each panel) are profiled with a free energy method with a per-target systematic error ($\sigma_{sys,ij,\ target}$) of 1 kcal/mol and some correlation coefficient (ρ). The top compounds that are predicted to have an improvement in selectivity greater than a set "synthesis rule" threshold (100×, 500×, or 1000×, Figure 1D, each curve) are synthesized, up to a maximum of 10 compounds. The expected speedup can then be calculated as the ratio of the number of synthesized compounds that have a true selectivity improvement of 2.8 kcal/mol (100× or 2 log units) to the number of compounds expected to have a true selectivity improvement of 2.8 kcal/mol had the same number of compounds as were synthesized been drawn randomly from the underlying unit normal distribution.

As shown in Figure 1D, the more stringent synthesis rules combined with high correlation coefficients (ρ) allow free energy calculations to have the highest impact in designing selectivity inhibitors, provided enough compounds have been scored. Interestingly, at correlation coefficient ρ=0.75 and low numbers of scored compounds, the 500× synthesis provides a greater speedup than 1000× synthesis rule. This is because there is high probability no compounds meet the more 1000× stringent synthesis rule until many more compounds are scored. This has implications for drug discovery efforts, where time and computational effort may limit the number of compounds able to be profiled with free energy methods.

An experimental data set of CDK2/CDK9 inhibitors demonstrates the difficulty in achieving high selectivity

To assess the correlation of errors in free energy predictions for selectivity, we set out to gather data sets that met a number of criteria. We searched for data sets that contained binding affinity data for a number of kinase targets and ligands in addition to crystal structures for each target with the same ligand.

This data set contains a congeneric series of ligands with experimental data for CDK2 and CDK9, with the

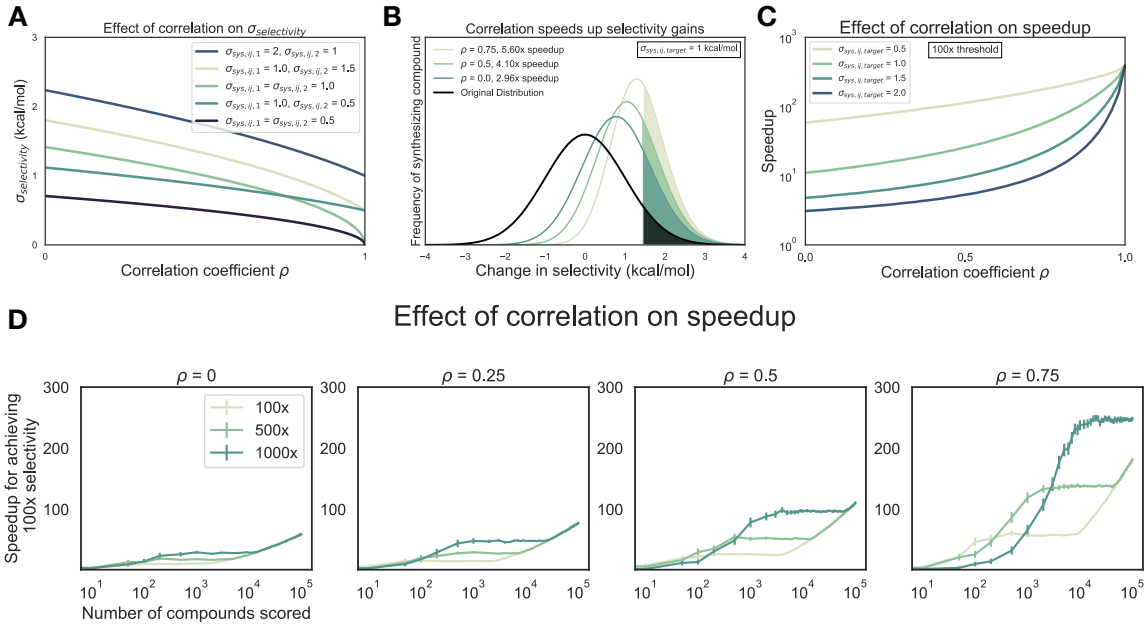


Figure 1. Free energy calculations can accelerate selectivity optimization. (**A**) The effect of correlation on expected errors for predicting selectivity ($\sigma_{selectivity}$) in kcal/mol when statistical error is negligible due to infinite sampling. Each curve represents a different combination of per target systematic errors ($\sigma_{sys,ij,1}$ and $\sigma_{sys,ij,2}$). (**B**) The change in selectivity for molecules proposed by medicinal chemists optimizing a lead candidate can be modeled by a normal distribution centered on zero with a standard deviation of 1 kcal/mol (black curve). Each green curve corresponds to the distribution of compounds made after screening for a 1 $\log_{10}$ unit (1.4 kcal/mol) improvement in selectivity with a free energy methodology with a 1 kcal/mol per target systematic error and a particular correlation, in the regime of infinite sampling where statistical error is zero. The shaded region of each curve corresponds to the compounds with a real 1 $\log_{10}$ unit improvement in selectivity. The speedup reflects the expected reduction in compounds that must be synthesized to reach a selectivity goal, and is calculated as the ratio of the percentage of compounds made with a real 1 $\log_{10}$ unit improvement to the percentage of compounds that would be expected in the original distribution. (**C**) The speedup (y-axis, log scale) expected for 100× (2 $\log_{10}$ units, or 2.8 kcal/mol) selectivity optimization as a function of correlation coefficient ρ. Each curve corresponds to a different value of $\sigma_{sys,ij,target}$. (**D**) The speedup (y-axis) expected for 100× (2 $\log_{10}$ units, or 2.8 kcal/mol) selectivity optimization as a function of number of compounds scored computationally (x-axis) and correlation coefficient ρ (each panel) for a method with per-target systematic error ($\sigma_{sys,ij,target}$) of 1 kcal/mol in the regime of infinite sampling. After profiling, the top compounds that meet or surpass the synthesis rule (the predicted selectivity threshold a compound must reach to be triggered for synthesis, each curve) are synthesized, up to a maximum of 10 synthesized compounds. Error bars (y-axis) represent the 95% CI for 1000 replicates at each point. The expected speedup is calculated as the ratio of the number of synthesized compounds that have a true selectivity improvement of 2.8 kcal/mol (100× or 2 log units) divided by the expectation of a compound showing a true selectivity improvement of 2.8 kcal/mol had the same number of compounds that were synthesized been drawn randomly from the underlying unit normal distribution. If no compounds were predicted to meet or surpass the synthesis rule, the speedup was assigned a default value of 1.

goal of potently inhibiting CDK9 and sparing CDK2. Based on a multiple sequence alignment of the 85 binding site residues identified in the kinase–ligand interaction fingerprints and structure (KLIFS) database [58, 59], CDK2 and CDK9 share 57% sequence identity (Supp. Table 1, Supp. Figure 1). For this CDK2/CDK9 data set [53], ligand 12c was cocrystallized with CDK2/cylin A (Figure 2A, left) and CDK9/cyclin T (Figure 2B, left), work that was published in a companion paper [60]. In both CDK2 and CDK9, ligand 12c forms relatively few hydrogen bond interactions with the kinase. Each kinase forms a pair of hydrogen bonds between the ligand scaffold and a hinge residue (C106 in CDK9 and L83 in CDK2) that is conserved across all of the ligands in this series. CDK9, which has slightly lower affinity for ligand 12c (Figure 2C, right), forms an interaction between the sulfonamide of ligand 12c and residue E107. On the other hand, CDK2 forms interactions between the sulfonamide of ligand 12c and residues K89 and H84. The congeneric series of ligands contains a number of difficult perturbations, particularly at substituent point R3 (Figure 2C, left). Ligand 12i also presented a challenging perturbation, moving the 1-(piperazine-1-yl)ethanone from the *meta* to *para* location.

This congeneric series of ligands also highlights two of the challenges of working from publicly available data: First, the dynamic range of selectivity is incredibly narrow, with a mean S (CDK9 - CDK2) of -0.65 kcal/mol, and a standard deviation of only 0.88 kcal/mol; the total dynamic range of this data set is 2.8 kcal/mol. Second, experimental uncertainties are not reported for the experimental measurements. This data set reported K_i values calculated from measured IC_{50}, using the K_m (ATP) for CDK2 and CDK9 and [ATP] from the assay using the Cheng-Prussof equations [61]. Thus, for this and subsequent sets of ligands, the random experimental uncertainty is assumed to be 0.3 kcal/mol based on previous work done to summarize uncertainty in experimental data, assuming there is no systematic experimental error. While K_i values are reported, these values are derived from IC50 measurements. A number of studies report on the reproducibility of intra-lab IC50 measurements. These values range from as low as 0.22 kcal/mol [62], from public data, to as high as 0.4 kcal/mol [6], which was estimated from internal data at Abbott Laboratories. The assumed value of 0.3 kcal/mol falls within this range, and agrees well with the uncertainty reported from Novartis for two different ligand series [63].

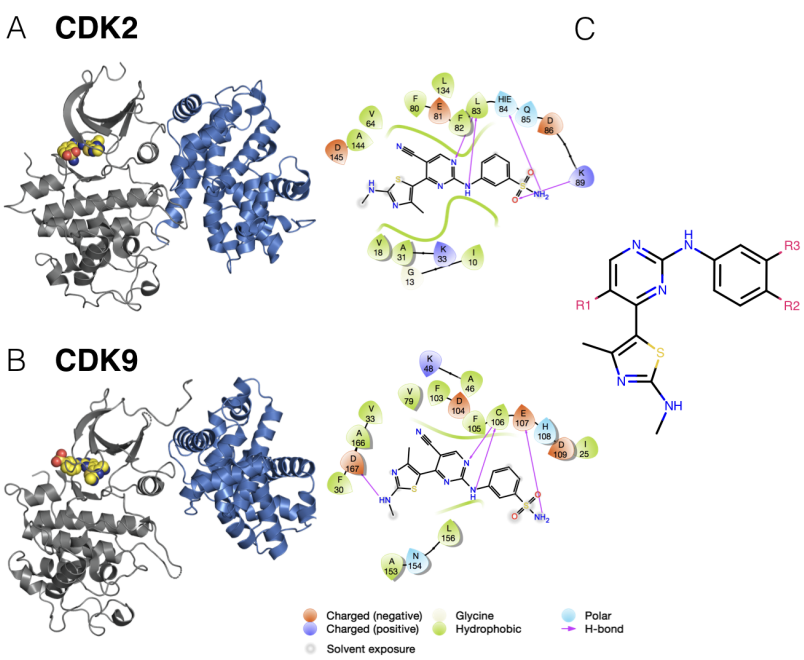
An experimental data set of CDK2/ERK2 inhibitors shows greater selectivity was achieved for a pair of more distantly related kinases

The CDK2/ERK2 data set from Blake et al. [54] also met the criteria described above, with the goal of developing a potent ERK2 inhibitor. Based on a multiple sequence alignment of the KLIFs binding site residues [58, 59], CDK2 and ERK2 share 52% sequence identity (Supp. Table 1, Supp. Figure 1), making them slightly less closely related than CDK2 and CDK9. While CDK2 and ERK2 both belong to the CMGC family of kinases, CDK2 is in the CDK subfamily, while ERK2 is in the MAPK subfamily.

Crystal structures for both CDK2 (Figure 3A, top) and ERK2 (Figure 3B, top) were available with ligand 22 (according to the manuscript numbering scheme) co-crystallized. Of note, CDK2 was not crystallized with cyclin A, despite cyclin A being included in the affinity assay reported in the paper [54].

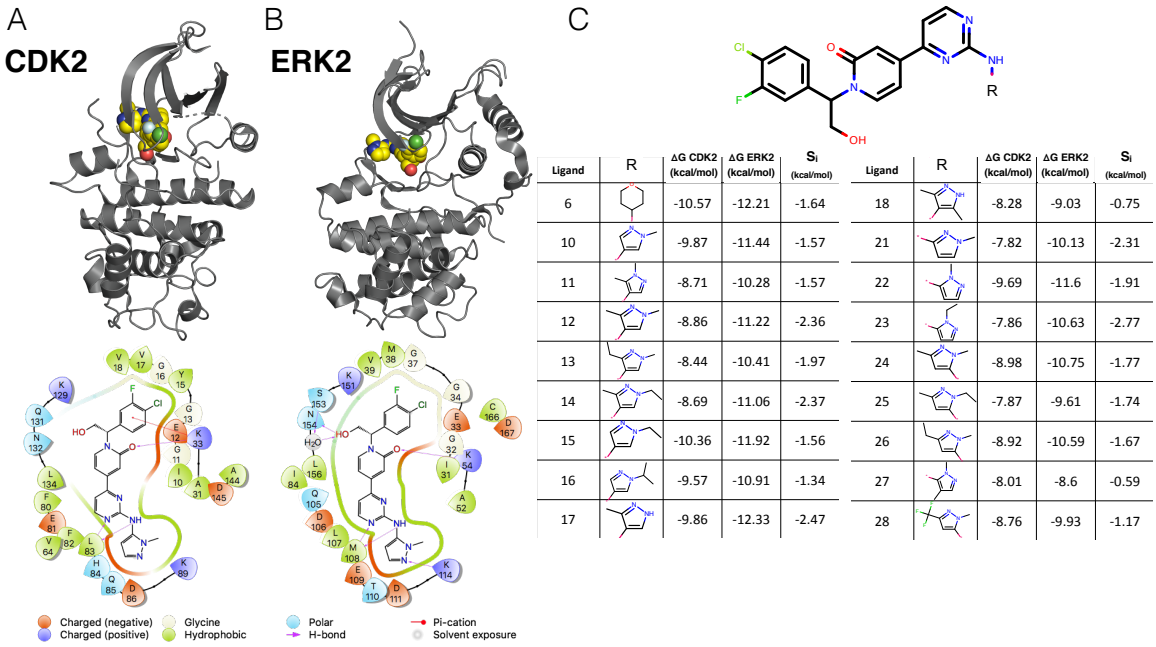
CDK2 in this crystal structure (4BCK) adopts a DFG-in conformation with the αC helix rotated out, away from the ATP binding site and breaking the conserved salt bridge between K33 and E51 (Supplementary Figure 2A), indicative of an inactive kinase [44, 64]. By comparison, the CDK2 structure from the CDK2/CDK9 data set adopts a DFG-in conformation with the αC helix rotated in, forming the ionic bond between K33 and E51 indicative of an active kinase, due to allosteric activation by cyclin A. While missing cyclins have caused problems for free energy calculations in prior work, it is possible that the fully active, cyclin-bound conformation contributes equally to binding affinity for all of the ligands in this series, and the high accuracy of the potency predictions (Figure 4, top left) is the result of fortuitous cancellation of errors.

The binding mode for this series is similar between both kinases. There is a set of conserved hydrogen bonds between the scaffold of the ligand and the backbone of one of the hinge residues (L83 for CDK2 and M108 for ERK2). The conserved lysine (K33 for CDK2 and K54 for ERK2), normally involved in the formation of a ionic bond with the αC helix, forms a hydrogen bond with the scaffold (Figure 3A and 3B, bottom) in both CDK2 and ERK2. However, in the ERK2 structure, the hydroxyl engages a crystallographic water as well as N154 in a hydrogen bond network that is not present in the CDK2 structure. The congeneric ligand series features a single solvent-exposed substituent. This helps to explain the narrow distribution of selectivities,



Ligand	R1	R2	R3	ΔG CDK2 (kcal/mol)	ΔG CDK9 (kcal/mol)	S_i (kcal/mol)
12a	CN	H		-12.27	-11.21	1.06
12b	OH	H		-7.23	-8.22	-0.99
12c	CN	H		-11.45	-11.21	0.24
12e	F	H		-11.62	-11.45	0.17
12f	Cl	H		-10.91	-10.85	0.06
12g	Methyl	H		-10.18	-11.32	-1.14
12h	Ethyl	H		-8.28	-9.56	-1.28
12j	CN	H		-10.04	-11.12	-1.08
12l	CN		H	-10.34	-10.44	-0.1
12n	CN	H		-10.06	-10.97	-0.91
12o	F	H		-10.06	-11.12	-1.06
12q	F	H		-10.91	-11.62	-0.71
12t	CN	H		-9.38	-11.12	-1.74
1a	H	H		-11.62	-11.86	-0.24
1b	H	H		-11.45	-11.86	-0.41

Figure 2. A CDK2/CDK9 data set illustrates selectivity optimization between closely-related kinases
Experimental IC_{50} data for a congeneric series of compounds binding to CDK2 and CDK9 was extracted from Shao et al. [53] and converted to free energies of binding. (**A**) *(left)* Crystal Structure (4BCK) [60] of CDK2 (gray ribbon) bound to ligand 12c (yellow spheres). Cyclin A is shown in blue ribbon. *(right)* 2D ligand interaction map of ligand 12c in the CDK2 binding site. (**B**) *(left)* Crystal structure of CDK9 (4BCI)[60] (gray ribbon) bound to ligand 12c (yellow spheres). Cyclin T is shown in blue ribbon. *(right)* 2D ligand interaction map of ligand 12c in the CDK9 binding site. (**C**) *(left)* 2D structure of the common scaffold for all ligands in congeneric ligand series 12 from the publication. *(right)* A table summarizing all R group substitutions as well as the published experimental binding affinities and selectivities [53], derived from the reported K_i as described in **Methods**.



Ligand	R	ΔG CDK2 (kcal/mol)	ΔG ERK2 (kcal/mol)	S_i (kcal/mol)	Ligand	R	ΔG CDK2 (kcal/mol)	ΔG ERK2 (kcal/mol)	S_i (kcal/mol)
6		-10.57	-12.21	-1.64	18		-8.28	-9.03	-0.75
10		-9.87	-11.44	-1.57	21		-7.82	-10.13	-2.31
11		-8.71	-10.28	-1.57	22		-9.69	-11.6	-1.91
12		-8.86	-11.22	-2.36	23		-7.86	-10.63	-2.77
13		-8.44	-10.41	-1.97	24		-8.98	-10.75	-1.77
14		-8.69	-11.06	-2.37	25		-7.87	-9.61	-1.74
15		-10.36	-11.92	-1.56	26		-8.92	-10.59	-1.67
16		-9.57	-10.91	-1.34	27		-8.01	-8.6	-0.59
17		-9.86	-12.33	-2.47	28		-8.76	-9.93	-1.17

Figure 3. A CDK2/ERK2 data set illustrates selectivity optimization among more distantly related kinases
(**A**) *(top)* Crystal structure of CDK2 (5K4J) shown in gray cartoon and ligand 22 shown in yellow spheres. *(bottom)* 2D interaction map of ligand 22 in the binding pocket of CDK2 (**B**) *(top)* Crystal structure of ERK2 (5K4I) shown in gray cartoon with ligand 22 shown in yellow spheres. *(bottom)* 2D interaction map of ligand 22 in the binding pocket of ERK2. (**C**) *(top)* Common scaffold for all of the ligands in the Blake data set [54], with R denoting attachment side for substitutions. *(bottom)* Table showing R group substitutions and experimentally measured binding affinities and selectivities, derived from the IC_{50} values as described in the methods section. Ligand numbers correspond to those used in the Blake publication [54].

with a mean selectivity of -1.74 kcal/mol (ERK2 - CDK2) and standard deviation of 0.56 kcal/mol; the total dynamic range of this data set is 2.2 kcal/mol. While the small standard deviation suggests that selectivity is difficult to drive with R-group substitution, the total dynamic range demonstrates that R-group substitutions can provide significant selectivity enhancements.

FEP+ calculations show smaller than expected errors for ΔS_{ij} predictions

Three replicates of FEP+ calculations were run on each target for both experimental data sets described above. The FEP+ predictions of the relative free energy of binding between ligands i and a reference compound (ref) for each target ($\Delta\Delta G_{i,\mathrm{ref},target}$) showed good accuracy and consistent results for all three replicates. The results for replicate 1 are reported in Figure 4 for both the CDK2 and ERK2 data set (bottom) and the CDK2/CDK9 data set (top), $\Delta\Delta G_{i,ref,\mathrm{target}}$ is defined for each ligand i using a consistent reference compound within data sets.

$$\Delta\Delta G_{i,\mathrm{ref,target}} = \Delta G_{i,\mathrm{target}} - \Delta G_{\mathrm{reference,\ target}} \tag{4}$$

The reference compounds (Compound 6 for CDK2/ERK2 and Compound 1a for CDK2/CDK9) were selected because they were the initial compounds from which the reported synthetic studies were started. Replicate 1 of the CDK2/ERK2 calculations is shown on the bottom of Figure 4, with an RMSE of $0.95^{1.25}_{0.63}$ and $0.97^{1.22}_{0.70}$ kcal/mol to CDK2 and ERK2, respectively. The RMSE reported here is calculated for all of the $\Delta\Delta G_{i,ref,\mathrm{target}}$ that were predicted. All of the CDK2 and ERK2 $\Delta\Delta G_{i,ref,\mathrm{target}}$s were predicted within 1 log unit of the experimental value. The change in selectivity (ΔS_{ij}) predictions show an RMSE of $1.41^{1.75}_{1.07}$ kcal/mol, with all the confidence intervals of the predictions falling within 1 log unit of the experimental values (Figure 4, top right panel). This was consistent across all three replicates of the calculations (Supp. Figure 6). Despite the low RMSE for the

selectivity predictions, the narrow dynamic range and high uncertainty from experiment and calculation makes it difficult to determine which compounds are more selective than others.

Replicate 1 of the CDK2/CDK9 calculations are shown in the top panel of Figure 4. The CDK2 and CDK9 data sets show higher errors in $\Delta\Delta G_{i,\text{ref,target}}$ predictions, with an RMSE of $1.15^{1.59}_{0.67}$ and $2.10^{2.65}_{1.47}$ kcal/mol respectively. This higher RMSE is driven by the reference compound, (Compound 1a) being poorly predicted, particularly in CDK9. There are a number of outliers that fall outside of 1 $\log_{10}$ unit from the experimental value for CDK9. While the higher per target errors make predicting potency more difficult, the selectivity predictions show a lower than expected RMSE of $1.37^{1.66}_{1.04}$ kcal/mol. This suggests that some correlation in the error is leading to fortuitous cancellation of the systematic error, leading to more accurate than expected predictions of ΔS_{ij}. These results were consistent across all three replicates of the calculation (Supp. Figure 4).

Correlation of systematic errors accelerates selectivity optimization

To quantify the correlation coefficient (ρ) of the systematic error between targets, we built a Bayesian graphical model to separate the systematic error from the statistical error and quantify our confidence in estimates of ρ (described in depth in Methods). Briefly, we modeled the absolute free energy (G) of each ligand in each thermodynamic phase (ligand-in-complex and ligand-in-solvent, with G determined up to an arbitrary additive constant for each phase) as in Equation 15. The model was chained to the FEP+ calculations by providing the $\Delta G^{calc}_{\text{phase},ij,\text{target}}$ for each edge from the FEP+ maps (where j is now not necessarily the reference compound) as observed data, as in Equation 17. As in Equation 19, the experimental data was modeled as a normal distribution centered around the true free energy of binding ($\Delta G^{true}_{i,\text{target}}$) corrupted by experimental error, which is assumed to be 0.3 kcal/mol from previous work done to quantify the uncertainty in publicly available data [6]. ΔG values derived from reported IC_{50}s or K_is, as described in the methods section, were treated as data observations (Equation 19) and the $\Delta G^{true}_{i,\text{target}}$ was assigned a weak normal prior (Equation 20).

The correlation coefficient ρ was calculated for each Bayesian sample from the model posterior according to equation 22. The CDK2/CDK9 calculations show strong evidence of correlation, with a correlation coefficient of $0.72^{0.83}_{0.58}$ (Figure 5A, right) for replicate 1. The rest of the replicates showed strong agreement (Supp. Figure 4). The joint marginal distribution of errors is strongly diagonal, which is expected based on the value for ρ (Figure 4A, left).

The joint marginal distribution of the error (ϵ) for each target is more diagonal than symmetric, which is expected for cases in which ρ is 0.4 (Supp. Figure 3). To quantify the expected speedup of selectivity with this level of correlation in the systematic errors for CDK2/CDK9, we first calculated the per target systematic error $\sigma_{\text{sys,ij,target}}$ by taking the mean of the absolute value of $\epsilon_{ij,\text{target}}$ where j is the reference compound 1a. Combining these estimates for the correlation coefficient (ρ) and the per target systematic errors ($\sigma_{\text{sys,ij,target}}$), we can compute $\sigma_{\text{selectivity}}$ and the expected speedup in the regime of infinite sampling effort where there is no statistical error when the number of compounds scored and synthesized is unlimited. The high correlation in errors for the CDK2/CDK9 calculations leads to a speedup of 3x for 1 $\log_{10}$ unit selectivity optimization and 10x for 2 $\log_{10}$ unit selectivity optimization (Figure 4A, right), despite the much high per target systematic errors ($\sigma_{\text{sys,ij,target}}$).

The correlation coefficient ρ for replicate 1 of the CDK2/ERK2 calculations was quantified to be $0.44^{0.70}_{0.12}$ (where the lower and upper values indicate a 95% confidence interval), indicating that the errors are moderately correlated between ERK2 and CDK2 (Figure 5B, right); this was consistent with the distribution for ρ in replicate 3 (Supp. Figure 7), while the confidence interval of ρ for replicate 2 is much wider, indicating the correlation is weak.

Considering the speedup model where the number of compounds scored and synthesized is unlimited, the modest correlation and low per target systematic errors for the CDK2/ERK2 calculations allow for a predicted 4–5x speedup for 1 $\log_{10}$ unit selectivity optimization, and a 30–40x speedup for 2 $\log_{10}$ unit selectivity optimization (Figure 5B, right).

Using the correlation coefficient (ρ), $\sigma_{\text{stat,ij,target}}$, and $\sigma_{\text{sys,ij,target}}$ quantified from the Bayesian model for each set of calculations, we can now calculate the y-axis error bars for the ΔS panels of Figure 4 according to the

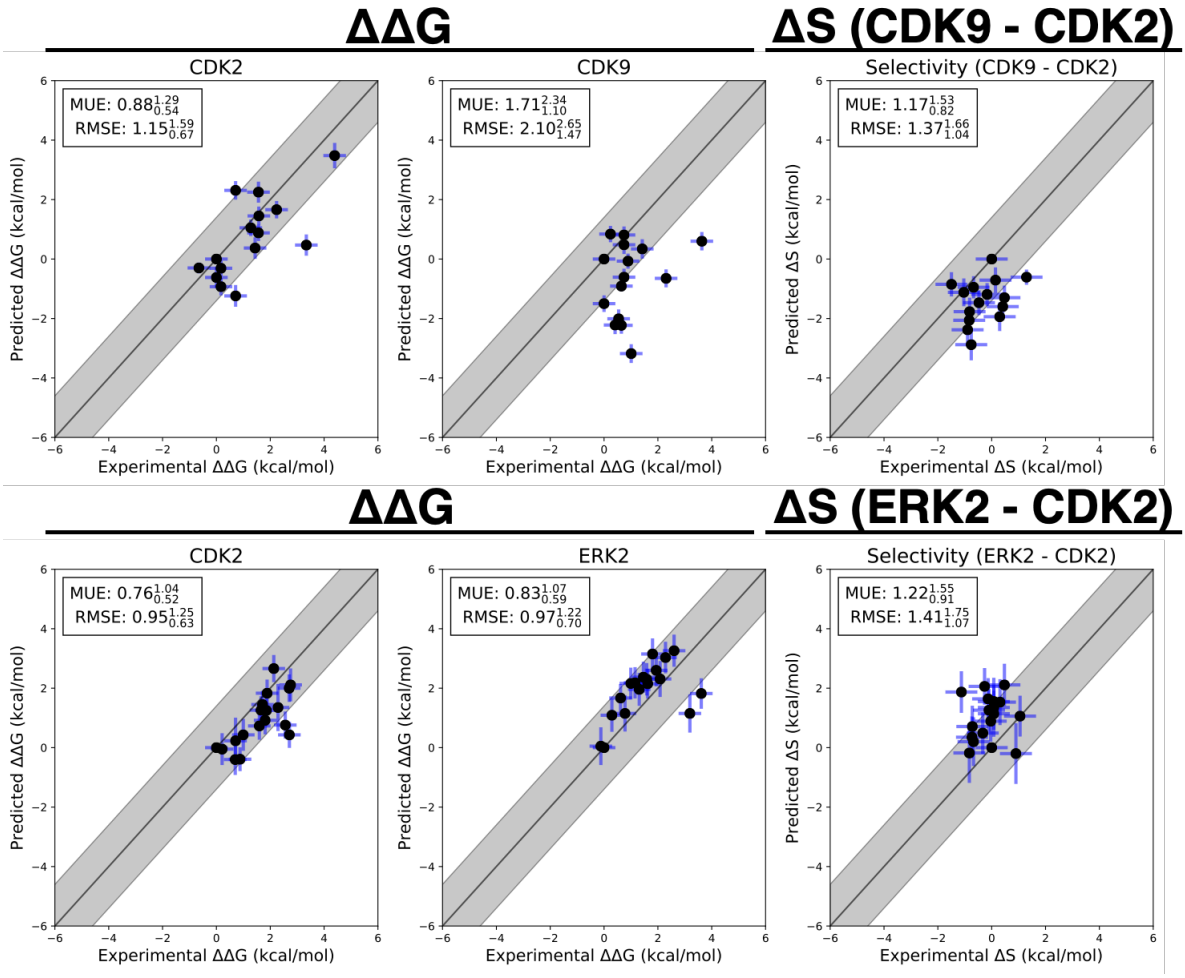


Figure 4. Selectivity predictions suggest correlation in systematic error
$\Delta\Delta G_{i,\text{ref,target}}$ and $\Delta S_{i,\text{ref}}$ predictions for CDK2/CDK9 (*top*) from the Shao data sets and CDK2/ERK2 from the Blake data sets (*bottom*). The experimental values are shown on the X-axis and calculated values on the Y-axis. Each data point corresponds to a transformation between a ligand i to a set reference ligand (ref) for a given target. All values are shown in units of kcal/mol. The horizontal error bars show to the $\delta\Delta\Delta G_{ij}^{exp}$ based on the assumed uncertainty of 0.3 kcal/mol[6, 63] for each ΔG_i^{exp}. We show the estimated statistical error ($\sigma_{\text{stat,ij,target}}$) as vertical blue error bars, which are one standard error. For selectivity, the errors were propagated under the assumption that they were completely uncorrelated. $\sigma_{\text{stat,ij,target}}$ was estimated by calculating the standard deviation of $\Delta\Delta G_{ij,\text{ target}}^{\text{FEP}}$ from the Bayesian model described in depth in **Methods**, where j is the reference compound. The black line indicates agreement between calculation and experiment, while the gray shaded region represent 1.36 kcal/mol (or 1 $\log_{10}$ unit) error. The mean unsigned error (MUE) and root-mean squared error (RMSE) are shown on each plot with bootstrapped 95% confidence intervals.

proposed $\sigma_{\text{selectivity}}$ equation (Eq 3). Shown in Supplemental Figure 9, we can see that $\sigma_{\text{selectivity}}$ accounts for most of the disagreement between the predicted ΔS_{ij} and the experimental ΔS_{ij}.

Expending more effort to reduce statistical error can be beneficial in selectivity optimization

Up to this point, we have considered only systematic error in quantifying the speedup free energy calculations can enable for selectivity optimization, by assuming enough sampling is done to reduce the statistical error for each target to zero. To begin understanding how statistical error impacts this speedup, we modified the model of speedup by additionally considering the per target statistical error ($\sigma_{\text{stat, target}}$), which we define in Equation 7 such that at the baseline effort, N, $\sigma_{\text{stat,ij,target}}$ is 0.2 kcal/mol. In this definition, it takes 4× the sampling, or effort, to reduce statistical error by a factor of 2×. We assume that statistical error is uncorrelated when propagating to two targets, and that $\sigma_{\text{sys,ij,target}}$ is ≈ 1.0 kcal/mol for both targets [4, 62]. As shown in Figure 6, expending effort to reduce $\sigma_{\text{stat,ij,target}}$ when ρ is less than 0.5 does not change the expected speedup for the 100× selectivity threshold in meaningful way, suggesting that it is not worth running calculations longer than the default protocol in this case. However, when $\rho > 0.5$, the curves do start to separate, particularly the 1/4×, 1×, and 4× effort curves. This suggests that when the correlation is high, running longer calculations can net improvements in selectivity optimization speed. Interestingly, the 16×, 48×, and ∞ effort curves do not differ greatly from the 4× effort curve, indicating that there are diminishing returns to running longer calculations.

The estimated correlation coefficient is robust to Bayesian model assumptions

In order to better understand the statistical error in our calculations, we performed three replicates of our calculations, and calculated the standard deviation of the cycle closure corrected $\Delta\Delta G$ for each edge of the map, and compared that value to the cycle closure errors and Bennett errors reported for each edge (Supp. Figure 8). For each set of calculations, the standard deviation suggests that the statistical error is between 0.1 and 0.3 kcal/mol, which is in good agreement with the reported Bennett error (Supp. Figure 8). However, hysteresis in the closed cycles in the FEP map as reflected by the cycle closure error estimates indicate much larger sampling errors than those estimated by the Bennett method or standard deviations of multiple runs, suggesting that both the Bennett errors and standard deviation of multiple replicates are underestimating the statistical error for these calculations. Based on this observation, we include a scaling parameter α in the Bayesian error model (Eq. 16) to account for the BAR errors underestimating the cycle closure statistical uncertainty. We also considered using a distribution with heavier tails, such as a Student's t-distribution, but found the quantification of the correlation coefficient ρ insensitive to the use of either a scaling parameter or heavier-tailed distributions (Supp. Fig. 10).

Discussion and Conclusions

S is a useful metric for selectivity in lead optimization

There are a number of different metrics for quantifying the selectivity of a compound [55], which look at selectivity from different views depending on the information trying to be conveyed. One of the earliest metrics was the standard selectivity score, which conveyed the number of inhibited kinase targets in a broad scale assay divided by the total number of kinases in the assay [65]. The Gini coefficient is a method that does not rely on any threshold, but is highly sensitive to experimental conditions because it is dependent on percent inhibition [66]. Other metrics take a thermodynamic approach to kinase selectivity and are suitable for smaller panel screens [67, 68]. Here, we propose a more granular, thermodynamic view of selectivity that is straightforward to calculate using free energy methods: the change in free energy of binding for a given ligand between two different targets (S). S is a useful metric of selectivity in lead optimization once a single, or small panel, of off-targets have been identified and the goal is to use physical modeling to either improve or maintain selectivity within a lead series.

Systematic error correlation can accelerate selectivity optimization

We have demonstrated, using a simple numerical model that assumes unlimited synthetic and computational resources, the impact that free energy calculations with even weakly correlated systematic errors can have on

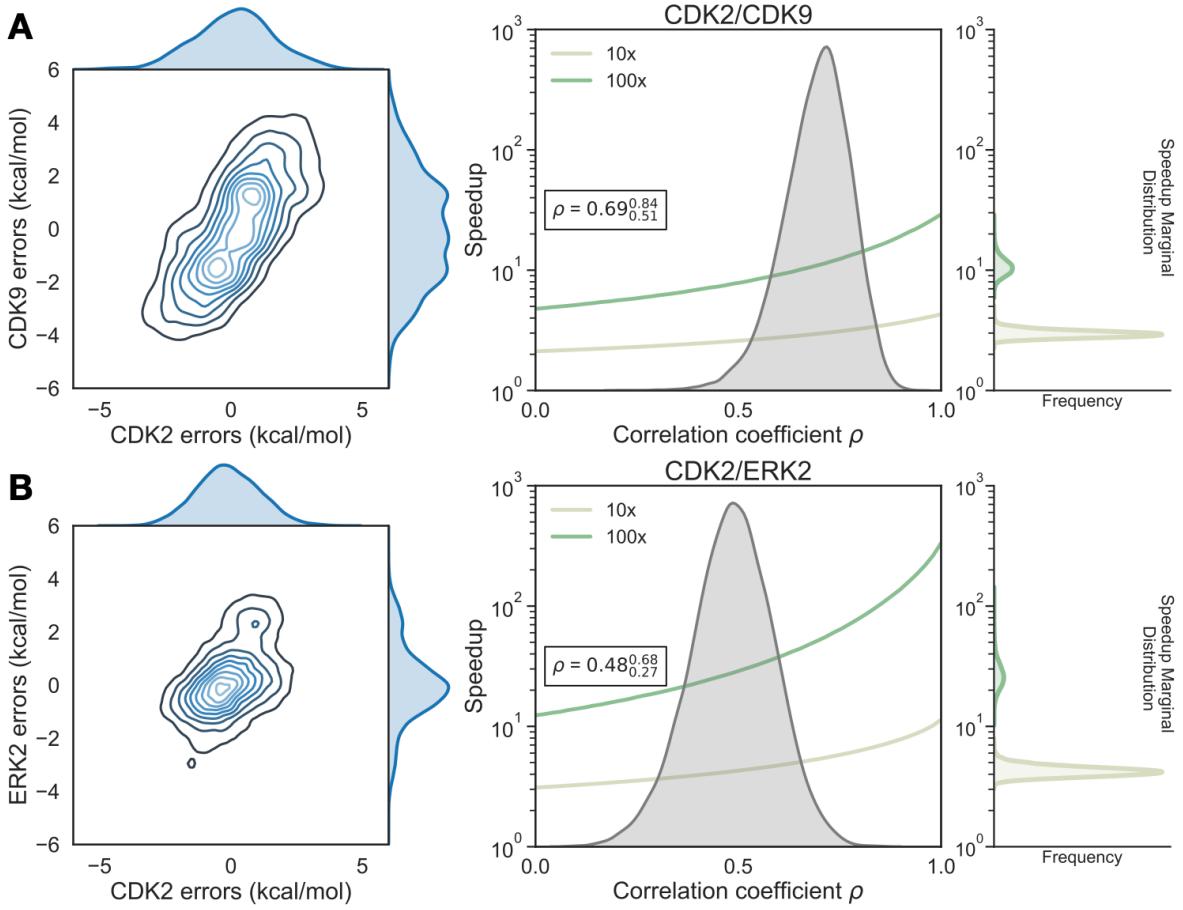


Figure 5. Correlation in systematic errors between targets can significantly accelerate selectivity optimization (**A**, *left*) The joint posterior distribution of the prediction errors for the more distantly related CDK2 (x-axis) and CDK9 (y-axis) from the Bayesian graphical model. (**A**, *right*) Speedup in selectivity optimization (y-axis), which estimates the reduction in compounds that must be synthesized to achieve a target selectivity when aided by free energy calculations, using the model where the number of compounds scored and synthesized is unlimited, as a function of correlation coefficient (x-axis). To calculate $\sigma_{\text{selectivity}}$, we calculate the per target systematic error ($\sigma_{\text{sys,ij,target}}$) by taking the mean of $\epsilon_{ij,\text{target}}$ where j is the reference compound 1a. The posterior marginal distribution of the correlation coefficient (ρ) is shown in gray, while the expected speedup is shown for 100× (green curve) and 10× (yellow curve) selectivity optimization. The inserted box shows the mean and 95% confidence interval for the correlation coefficient. The marginal distribution of speedup is shown on the right side of the plot for both 100× (green) and 10× (yellow) selectivity optimization speedups. (**B**) As above, but for the more closely related CDK2/ERK2 selectivity data set using compound 6 as the reference.

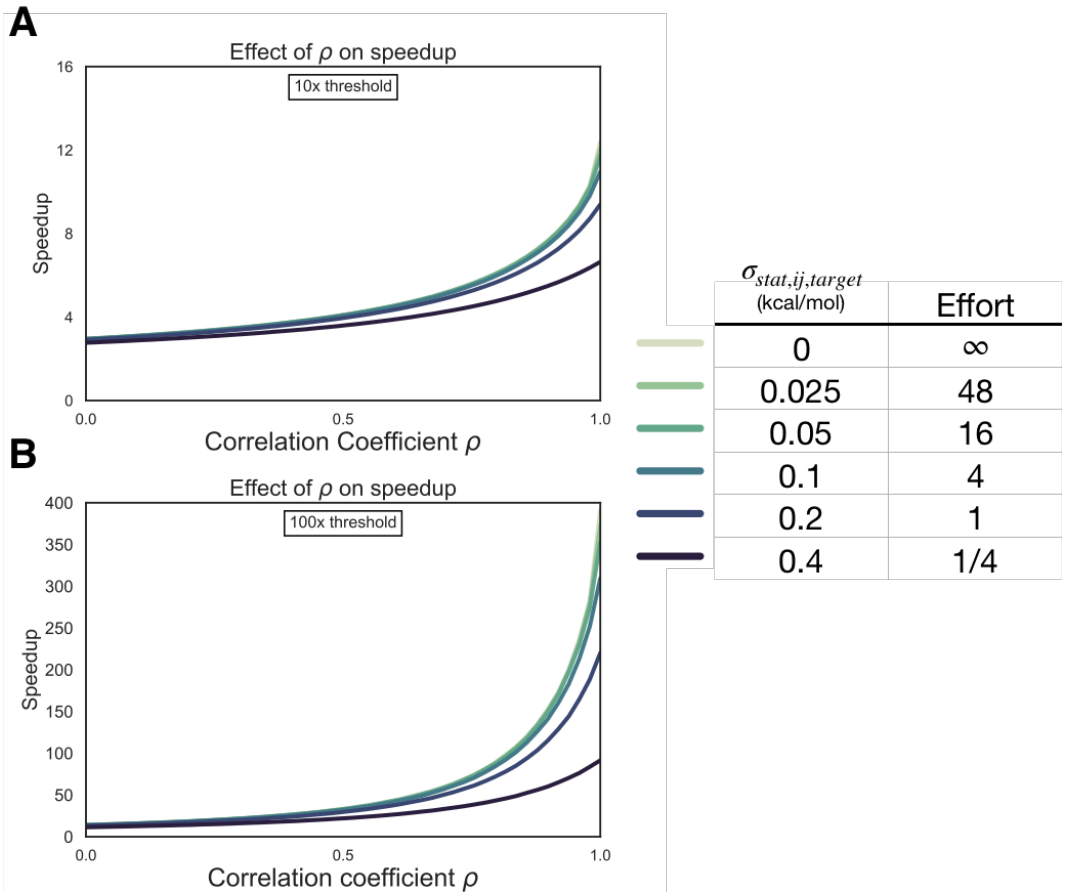


Figure 6. Reducing statistical uncertainty when systematic error correlation is high improves the speedup in selectivity optimization achievable with free energy calculations. (*left*) The speedup in selectivity (Y-axis) as a function of correlation coefficient (X-axis). Each curve represents a different per target statistical error ($\sigma_{stat,ij,target}$) for 10× (1 $\log_{10}$ unit) (**A**) and 100× (2 $\log_{10}$ unit) (**B**) thresholds (*right*) Table with the per target statistical error ($\sigma_{stat,ij,target}$), kcal/mol) corresponding to each curve on the left and a rough estimate of the generic amount of computational effort it would take to achieve that statistical uncertainty.

speeding up the optimization of selectivity in small molecule kinase inhibitors. While the expected speedup is dependent on the per target systematic error of the method ($\sigma_{sys,ij,target}$), the speedup is also highly dependent on the correlation of errors made for both targets. Unsurprisingly, free energy methods have greater impact as the threshold for selectivity optimization goes from 10× to 100×. While 100× selectivity optimization is difficult to achieve, the expected benefit from free energy calculations is also quite high, with speedups of one or two orders of magnitude possible. In a more realistic scenario, where the number of compounds scored and synthesized is limited by resources, we have demonstrated using the same numerical model that more stringent synthesis rules results in increased speedup from free energy calculations. This holds true across different correlation coefficients (ρ), provided enough compounds are scored. As our model shows, it is possible for stringent synthesis rules to provide benefits similar to operating with high systematic error correlation coefficients (ρ).

Two pairs of kinase test systems suggest systematic errors can be correlated

To quantify the correlation of errors in two example systems, we gathered experimental data for two congeneric ligand series with experimental data for CDK2 and ERK2, as well as CDK2 and CDK9. These data sets, which had crystal structures for both targets with the same ligand co-crystallized, exemplify the difficulty in predicting selectivity. The dynamic range of selectivity for both systems is very narrow, with most of the perturbations not having a major impact on the overall selectivity achieved. Further, the data was reported without reliable experimental uncertainties, which makes quantifying the errors made by the free energy calculations difficult. This issue is common when considering selectivity, as many kinase-oriented high throughput screens are carried out at a single concentration and not highly quantitative.

The CDK9 calculations contained an outlier, compound 12h, that drove much of the prediction error for that set. Compound 12e (R1 = F) is the most potent against CDK9 of the compounds in with a sulfonamide at R3 (Figure 2). The addition of a single methyl group decreases the potency against CDK9 (compound 12g) and while only slightly changing the affinity for CDK2. However, adding on another methyl group (compound 12h) results in an order of magnitude decrease in K_i for both CDK9 and CDK2. Crystal structures for both kinases show that R1 points into a pocket formed by the backbone, and the sidechains of a Valine and Phenylalanine. While ethyl at R1 in compound 12h *is* bulkier, the magnitude of the decrease in affinity for both kinases is larger than might be expected, given that the pocket suggests an ethyl group would be well accommodated in terms of fit and the hydrophobicity of the sidechains. For both kinases, the free energy calculations predict that this addition should *improve* the potency, suggesting that it is possible that the model is missing a chemical detail that might explain the trend seen in the experimental data. We expect that these types of errors, which would be troubling when predicting potency alone, will drive the correlation of systematic errors and fortuitously cancel when predicting selectivity.

Despite CDK2 and ERK2 being more distantly related than CDK2 and CDK9, the calculated correlation in the systematic error for two of the replicates suggests that fortuitous cancellation of errors may be applicable in a wider range of scenarios than closely related kinases within the same family. However, the confidence interval of the correlation is quite broad, including 0 for the lower bound for the third replicate, suggesting that errors for more distantly related proteins will have only moderate, if any, correlation.

Reducing statistical error is beneficial when systematic errors are correlated

In order to better understand if there are situations where it is beneficial to run longer calculations to minimize statistical error to achieve a larger speedup in the synthesis of selective compounds, we built a numerical model of the impact of statistical error in the context of different levels of systematic error correlation. Our results suggest that unless the correlation coefficient $\rho > 0.5$ for the two targets of interest, there is not much benefit in running longer calculations. However, when the systematic error is reduced by correlation, longer calculations can help realize large increases in speedup to achieve selectivity goals. Keeping a running quantification of ρ for free energy calculations as compounds are made and the predictions can be tested will allow for decisions to be made about whether running longer calculations is worthwhile. It will also allow for an estimate of $\sigma_{selectivity}$, which is useful for estimating expected systematic error for prospective predictions. Importantly, we expect that correlation will be modeling protocol dependent and

any changes to the way the system is modeled over the course of discovery program are expected to change the observed correlation in the systematic error.

Larger data sets with a wide range of protein targets will enable future work

The data sets gathered here were limited by the total number of compounds, the small dynamic range for selectivity (S), and the lack of reliable experimental uncertainties. The small size of the data set makes it difficult to draw broad conclusions about the correlation in systematic errors. Understanding the degree of correlation *a priori* based on structural or sequence similarity requires study on a larger range of targets than the two pairs presented in this study. A larger data set that contained many protein targets, crystal structures, and quantitative binding affinity data would be ideal to draw conclusions about the broader prevalence of systematic error correlation.

This work demonstrates that correlation in the systematic errors can allow free energy calculations to facilitate significant speedups in selectivity optimization for drug discovery projects. This is particularly important in kinase systems, where considering multiple targets is an important part of the development process. The results suggest that free energy calculations can be particularly helpful in the design of kinase polypharmacological agents, especially in cases where there is high correlation in the systematic errors between multiple targets.

Methods

Numerical model of selectivity optimization speedup

To model the impact correlation of systematic error would have on the expected uncertainty for selectivity predictions, $\sigma_{\text{selectivity}}$ was calculated using Equation 3 for 1000 evenly spaced values of the correlation coefficient (ρ) from 0 to 1, for a number of combinations of per target systematic errors ($\sigma_{\text{sys,ij,1}}$ and $\sigma_{\text{sys,ij,2}}$). In the regime of infinite sampling and zero statistical error, the second term reduces to zero.

$$\sigma_{\text{selectivity}} = \sqrt{\sigma^2_{\text{sys,ij,1}} + \sigma^2_{\text{sys,ij,2}} - 2\rho\,\sigma_{\text{sys,ij,1}}\,\sigma_{\text{sys,ij,2}} + \sigma^2_{\text{stat,ij,1}} + \sigma^2_{\text{stat,ij,2}}} \tag{3}$$

The speedup in selectivity optimization that could be expected from using free energy calculations of a particular per target systematic error ($\sigma_{\text{sys,ij,target}}$) was quantified as follows using NumPy (v 1.14.2). An original, true distribution for the change in selectivity of 200 000 000 new compounds proposed with respect to a reference compound was modeled as a normal distribution centered around 0 with a standard deviation of 1 kcal/mol. This assumption was made on the basis that the majority of selectivity is driven by the scaffold, and R group modifications will do little to drive changes in selectivity. The 1 kcal/mol distribution is supported by the standard deviations of the selectivity in the experimental data sets referenced in this work, which are all less than, but close to, 1 kcal/mol.

In this model, we suppose that each of proposed compound is triaged by a free energy calculation and only proposed compounds predicted to increase selectivity by $\Delta S_{\text{ij}} \geq$1.4 kcal/mol (1 $\log_{10}$ unit) with respect to a reference compound would be synthesized. Based on reported estimates in the literature, we presume that relative free energy calculations have a per-target systematic error $\sigma_{\text{sys,ij,target}} \approx$1 kcal/mol [4], and explore the impact of the correlation coefficient ρ governing the correlation of these predictions between targets. The standard error in predicted selectivity, $\sigma_{\text{selectivity}}$, is given by Equation 3. When sampling is infinite and $\sigma_{\text{stat,ij,target}}$ is zero, $\sigma_{\text{selectivity}}$ is driven entirely by the systematic error component ($\sigma_{\text{sys,ij,target}}$), resulting in the error in predicted change in selectivity ΔS_{ij} modeled as a normal distribution centered around 0 with a standard deviation of $\sigma_{\text{sys,ij,target}}$ and added to the "true" ΔS_{ij},

$$\Delta S_{\text{ij, predicted}} = \Delta S_{\text{ij,true}}\left(\mathcal{N}_{\text{true}}(\mu = 0, \sigma^2 = 1)\right) + \Delta S_{\text{systematic error}}\left(\mathcal{N}_{\text{error}}(\mu = 0, \sigma^2_{\text{sys,ij,target}}(\rho)\right) \tag{5}$$

We ignore the potential complication of finite experimental error in this thought experiment, presuming the experimental uncertainty is sufficiently small as to be negligible.

The *speedup* in synthesizing molecules that reach this 10× selectivity gain threshold is calculated, as a function of ρ, as the ratio of the number of compounds that exceed the selectivity threshold in the case that

molecules predicted to fall below this threshold by free energy calculations were triaged and not synthesized, divided by the number of compounds that exceeded the selectivity threshold without the benefit of free energy triage. This process was repeated for a 100× (2.8 kcal/mol, 2 $\log_{10}$ unit) selectivity optimization and 50 linearly spaced values of the correlation coefficient (ρ) between 0 and 1, for four values of $\sigma_{\text{selectivity}}$, using a sample size of 4×10^7 compounds.

The above model assumes that the number of compounds scored and synthesized is essentially unlimited. To assess the impact these methods might have on real drug discovery projects, where the number of compounds scored and synthesized are limited by computational and chemistry resources, we altered the above model to consider the number of compounds scored, the number of compounds triggered for synthesis, and the threshold a compound needed to reach in order to be considered for synthesis.

We repeated the mode detailed above, this time scoring only the following numbers of compounds: 10, 50, 100, 200, 500, the range from 1000 to 10000 in steps of 1000, and the range from 10000 to 100 000 in steps of 2000. Compounds were drawn from a true distribution of $\Delta S_{ij,true}$ $\left(\mathcal{N}_{\text{true}}(\mu = 0, \sigma^2 = 1)\right)$ and triaged using a free energy method as detailed above with a per-target systematic error ($\sigma_{\text{sys,ij,target}}$) of 1 kcal/mol. The top predicted compounds that meet or surpass a synthesis rule, up to a maximum of 10 compounds, are selected for synthesis. Here, we consider synthesis rules of 100×, 500× and 1000× when trying to design 100× (2.8 kcal/mol, 2 $\log_{10}$ unit) improvements in selectivity. The *speedup* was calculated as the number of synthesized compounds whose $\Delta S_{ij,true}$ reaches the desired 100× threshold divided by the expected value ($\mathbb{E}_{selective}$) for a selective compound given the number of synthesized compounds. This expectation can be calculated as,

$$\mathbb{E}_{selective} = P(\Delta S_{ij} > \text{threshold} \mid \mathcal{N}_{true}) * n_{\text{synthesized}} \tag{6}$$

Where $P(\Delta S_{ij} > \text{threshold} \mid \mathcal{N}_{true})$ is the probability $\Delta S_{ij,true}$ for some compound is better than a particular selectivity threshold given the distribution of $\Delta S_{ij,true}$ $\left(\mathcal{N}_{true}(\mu = 0, \sigma^2 = 1)\right)$ for 100 000 000 compounds, and $n_{\text{synthesized}}$ is the number of compounds synthesized. If no compounds were predicted to meet or surpass the synthesis rule, the speedup was assigned a default value of 1. We performed 1000 replicates of each condition and report the mean and 95 % CI in Figure1D.

Numerical model of impact of statistical error on selectivity optimization

To model the impact of finite statistical error in the alchemical free energy calculations, a similar scheme was used with the following modifications: Each proposed compound was triaged by a free energy calculation with a per target systematic error ($\sigma_{\text{sys,ij,target}}$) of 1.0 kcal/mol [4] and a specified correlation coefficient ρ. A $\sigma_{\text{selectivity}}$ was calculated according to Equation 3, this time considering the statistical terms as non-negligible. The per target statistical error ($\sigma_{\text{stat,ij,target}}$) was defined as,

$$\sigma_{\text{stat,ij,target}} = \frac{\sigma_{\text{stat,base}}}{\sqrt{N}} \tag{7}$$

where N is the relative effort put into running sampling the calculation and $\sigma_{\text{stat,base}}$ is such that when $N = 1$, $\sigma_{\text{stat,ij,target}}$ = 0.2 kcal/mol. The statistical error is propagated assuming it is uncorrelated, as independent sets of calculations are used for each target, giving us the second set of terms in 3. This gives an updated model for the error in predicted change in selectivity ΔS_{ij}. The systematic and statistical errors were modeled as Gaussian noise added to the true distribution,

$$\begin{aligned}\Delta S_{\text{ij,predicted}} = \Delta S_{\text{ij,true}}\left(\mathcal{N}_{\text{true}}(\mu = 0, \sigma^2 = 1)\right) &+ \Delta S_{\text{systematic error}}\left(\mathcal{N}_{\text{systematic}}(\mu = 0, \sigma^2_{\text{sys,ij,target}}(\rho))\right) \\ &+ \Delta S_{\text{statistical error}}\left(\mathcal{N}_{\text{statistical}}(\mu = 0, \sigma^2_{\text{stat,ij,target}})\right)\end{aligned} \tag{8}$$

Any compound predicted to have an improvement in selectivity of above the threshold (either 1.4 kcal/mol (1 $\log_{10}$ units) or 2.8 kcal/mol (2 $\log_{10}$ units)) would then be made and have its selectivity experimentally measured, using an experimental method with perfect accuracy. The speedup value for each value of ρ is calculated as previously described.

Binding Site Similarity analysis

To quantify the similarity between the different kinase pairs, a structure-informed binding site sequence comparison was performed. In the KLIFS database, the binding site of typical human kinases is defined by 85 residues, comprising known kinase motives (DFG, hinge, G-loop, aC-helix, ...), which cover potential interactions with type I-IV inhibitors [58, 59]. KLIFS provides a multiple sequence alignment in which each kinase sequence is mapped to these 85 binding site residues. This mapping was used to calculate the sequence identify between the three kinases CDK2, CDK9, and ERK2 used in this study (Supp. Figure 1 and Supp. Table 1). The score shows the percentage of identical residues between two kinases with respect to the 85 positions.

Extracting the binding free energy ΔG from reported experimental data

K_i values were derived from IC_{50} measurements reported for the ERK2/CDK2 data set (Figure 3), assuming Michaelis-Menten binding kinetics for an ATP-competitive inhibitor,

$$\mathrm{K}_i = \frac{IC_{50}}{1 + \frac{[S_0]}{K_m}} \tag{9}$$

Where the Michaelis-Menten constant for ATP (K_m (ATP)) is much larger than the initial concentration of ATP, S_0, so that $\mathrm{IC}_{50} \approx \mathrm{K}_i$.

These K_i values were then used to calculate a ΔG (Equation 10),

$$\Delta G = -k_B T \ln K_i \tag{10}$$

Here, k_B is the Boltzmann constant and T is absolute temperature (taken to be room temperature, $T \sim 300$K).

For the CDK2/CDK9 data set, the authors note that the assumption K_m (ATP) $\gg \mathrm{S}_0$ does *not* hold, and report K_is derived from their IC_{50} measurements using the K_m (ATP) for each kinase, as well as the S_0 from their assay. These values were then converted to ΔG using Equation 10. For both data sets, these derived ΔG were used to calculate $\Delta\Delta G$ between ligands for each kinase target.

As mentioned above, the assumption that K_m (ATP) $\gg \mathrm{S}_0$ may not always hold, and changes in IC_{50} may be driven by factors other than changes in ligand binding affinity. However, these assumptions have been used successfully to estimate relative free energies previously [62, 69]. Further, data was taken from the same lab and assay for each target. By using assays with the same kinase construct and ATP concentration, the relative free energies ($\Delta\Delta G_{ij}$) should be well determined for compounds within the assay. Even if the absolute free energies (ΔG_i) are off due to uncertainties in K_m (ATP) or S_0, they will be off by the same constant, which will cancel when calculating $\Delta\Delta G_{ij}$.

Structure Preparation

Structures from the Shao [53] (CDK2/CDK9), Hole [60] (CDK2/CDK9), and Blake [54] (CDK2/ERK2) papers were downloaded from the PDB [70], selecting structures with the same co-ligand crystallized.

For the Shao (CDK2/CDK9) data set, PDB IDs 4BCK (CDK2) and 4BCI (CDK9) were selected, which have ligand 12c cocrystallized. For the Blake data set (ERK2/CDK2), 5K4J (CDK2) and 5K4I (ERK2) were selected, cocrystallized with ligand 21. The structures were prepared using Schrodinger's Protein Preparation Wizard [71] (Maestro, Release 2017-3). This pipeline modeled in internal loops and missing atoms, added hydrogens at the reported experimental pH (7.0 for the Shao data set, 7.3 for the Blake data set) for both the protein and the ligand. All crystal waters were retained. The ligand was assigned protonation and tautomer states using Epik at the experimental pH±2, and hydrogen bonding was optimized using PROPKA at the experimental pH±2. Finally, the entire structure was minimized using OPLS3 with an RMSD cutoff of 0.3Å.

Ligand Pose Generation

Ligands were extracted from the publication entries in the BindingDB as 2D SMILES strings. 3D conformations were generated using LigPrep with OPLS3 [4]. Ionization state was assigned using Epik at experimental pH±2. Stereoisomers were computed by retaining any specified chiralities and varying the rest. The tautomer and ionization state with the lowest Epik state penalty was selected for use in the calculation. Any ligands predicted to have a positive or negative charge in its lowest Epik state penalty was excluded, with the

exception of Compound 9 from the Blake data set. This ligand was predicted to have a +1 charge for its lowest state penalty state. The neutral form the ligand was include for the sake of cycle closure in the FEP+ map, but was ignored for the sake any analysis afterwards. Ligand poses were generated by first aligning to the co-crystal ligand using the Largest Common Bemis-Murcko scaffold with fuzzy matching (Maestro, Release 2017-3). Ligands that were poorly aligned or failed to align were then aligned using Maximum Common Substructure (MCSS). Finally, large R-groups conformaitons were sampled with MM-GBSA using a common core restraint, VSGB solvation model, and OPLS3 force field. No flexible residues were defined for the ligand.

Free Energy Calculations

The FEP+ panel (Maestro, Release 2017-3) was used to generate perturbation maps. FEP+ calculations were run using the FEP+ panel from Maestro release 2018-3 in order to take advantage of the newest force field (OPLS3e) parameters available at the time. Any missing ligand torsions were fit using the automated FFbuilder protocol [7]. Custom charges were assigned using the OPLS3e force field using input geometries, according to the automated FEP+ workflow in Maestro Release 2018-3. Neutral perturbations were run for 15 ns per replica, using an NPT ensemble and water buffer size of 5Å. The SPC water model was used. A GCMC solvation protocol was used to sample buried water molecules in the binding pocket prior to the calculation, which discards any retained crystal waters.

Statistical Analysis of FEP+ calculations

To quantify the overall errors in the FEP+ calculations, we computed the mean unsigned error (MUE),

$$\mathrm{MUE} = \frac{\sum_0^n \mid \Delta\Delta G^{\mathrm{calc}}_{\mathrm{i,ref,target}} - \Delta\Delta G^{\mathrm{exp}}_{\mathrm{i,ref,target}} \mid}{n} \tag{11}$$

and the root mean squared error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_0^n (\Delta\Delta\, G^{\mathrm{calc}}_{\mathrm{i,ref,target}} - \Delta\Delta\, G^{\mathrm{exp}}_{\mathrm{i,ref,target}})^2}{n}} \tag{12}$$

MUE and RMSE were computed for $\Delta\Delta G_{ij,\mathrm{target}}$. For each ligand i, $\Delta\Delta G_{i,ref,\mathrm{target}}$ is defined where ref is a reference compound.

$$\Delta\Delta G_{i,\mathrm{ref,target}} = \Delta G_{i,\mathrm{target}} - \Delta G_{\mathrm{reference,\ target}} \tag{13}$$

For the CDK2/CDK9 data set, compound 1a was used as the reference compound, as it was the first compound from which the others in the series were derived. For the CDK2/ERK2 data set, compound 6 was used as the reference compound, since it was the compound from which the investigation was launch. A metabolite of compound 6 (not included in the data set here) was used as the starting compound from which the rest were derived. To account for the finite ligand sample size, we used 10 000 replicates of bootstrapping with replacement to estimate 95% confidence intervals. The code used to bootstrap these values is available on GitHub [https://github.com/choderalab/selectivity].

To compute the per-target statistical error ($\sigma_{\mathrm{stat,ij,target}}$) for each i,ref pair of ligands, we used the standard deviation of $\Delta\Delta G^{\mathrm{FEP}}_{ij,\ \mathrm{target}}$, where j is the reference compound, from the Bayesian model described in depth below in the **Methods** section. To compute the per target systematic error ($\sigma_{\mathrm{sys,ij,target}}$), we calculated the mean of $\epsilon_{ij,\mathrm{target}}$, where j is the reference compound, described in equation 21 in the Bayesian Model section of the **Methods**.

Quantification of the correlation coefficient ρ

To quantify ρ, we built a Bayesian graphical model using `pymc3` 3.5 [72] and theano 1.0.3 [73]. All code for this model is available on GitHub [https://github.com/choderalab/selectivity].

For each phase (complex and solvent), the prior for the absolute free energy (G) of ligand i (up to an arbitrary additive constant for each thermodynamic phase, ligand-in-complex or ligand-in-solvent), was treated as a normal distribution (Equation 15).

$$G^{phase}_{i,\mathrm{target}} \sim \mathcal{N}(\mu = 0,\ \sigma = 25.0\ \mathrm{kcal/mol}) \tag{14}$$

To improve sampling efficiency, for each phase, one ligand was chosen as the reference, and pinned to an absolute free energy of $G = 0$, with a standard deviation of 1 kcal/mol.

$$G^{phase}_{1,\text{target}} \sim \mathcal{N}(\mu = 0,\ \sigma = 1.0 \text{ kcal/mol}) \tag{15}$$

For each edge of the FEP map (ligand i –> ligand j), there is a contribution from dummy atoms, that was modeled as in Equation 16. Note that here, unlike what was done in Figure 4, ligand j is not necessarily a reference compound.

$$c_{i,j} \sim \mathcal{N}(\mu = 0,\ \sigma = 25.0 \text{ kcal/mol}) \tag{16}$$

The model was conditioned by including data from the FEP+ calculation.

$$\Delta G^{\text{calc}}_{\text{phase},\ ij,\ \text{target}} \sim \mathcal{N}(G^{\text{phase}}_{j,\text{target}} - G^{\text{phase}}_{i,\text{target}},\ \alpha\delta^2\Delta G^{\text{BAR}}_{\text{phase},\ ij,\ \text{target}}) \tag{17}$$

where $\delta^2\Delta G^{\text{BAR}}_{\text{phase},\ ij,\ \text{target}}$ is the reported BAR uncertainty from the calculation, and $\Delta G^{\text{calc}}_{\text{phase},\ ij,\ \text{target}}$ is the BAR estimate of the free energy for the perturbation between ligands i and j in a given phase. α is a scaling parameter shared by all $\Delta G^{\text{calc}}_{\text{phase},\ ij,\ \text{target}}$ for each target. Such scaling is necessary to account for the BAR statistical uncertainty underestimating cycle closure statistical uncertainty of our calculations, shown by Supp. Figure 8.

From this, we can calculate the $\Delta G^{\text{FEP}}_{i,\ \text{target}}$ for each ligand and target,

$$\Delta G^{\text{FEP}}_{i,\ \text{target}} = G^{\text{complex}}_{i,\text{target}} - G^{\text{solvent}}_{i,\text{target}} \tag{18}$$

From $\Delta G^{\text{FEP}}_{i,\ \text{target}}$, we calculated $\Delta\Delta G^{\text{FEP}}_{ij,\ \text{target}}$ for each pair of ligands, filtering out pairs where i and j are the same ligand and where the reciprocal was already included.

The experimental binding affinity was treated as a true value ($\Delta G^{\text{true}}_{i,\text{target}}$) corrupted by experimental uncertainty, which is assumed to be 0.3 kcal/mol [6]. There are a number of studies that report on the reproducibility and uncertainty of intra-lab IC_{50} measurements, ranging from as small as 0.22 kcal/mol [62] to as high as 0.4 kcal/mol [6]. The assumed value falls within this range and is in good agreement with the uncertainty reported from multiple replicate measurements in internal data sets at Novartis [63].

The values reported in the papers ($\Delta G^{\text{obs}}_{i,\text{target}}$) were treated as observations from this distribution (Equation 19),

$$\Delta G^{\text{obs}}_{i,\text{target}} \sim \mathcal{N}(\mu = \Delta G^{true}_{i,\text{target}},\ \sigma = 0.3 \text{ kcal/mol}) \tag{19}$$

$\Delta G^{\text{true}}_{i,\text{target}}$ was assigned a weak normal prior, as in Equation 20,

$$\Delta G^{\text{true}}_{i,\text{target}} = \mathcal{N}(\mu = 0, \sigma = 50 \text{ kcal/mol}) \tag{20}$$

$\Delta\Delta G^{\text{true}}_{ij,\ \text{target}}$ for each pair of ligands was calculated from $\Delta G^{\text{true}}_{i,\text{target}}$, filtering out pairs where i and j are the same ligand and where the reciprocal was already included as above.

The error for a given ligand was calculated as

$$\epsilon_{ij,\text{target}} = \Delta\Delta G^{\text{FEP}}_{ij,\ \text{target}} - \Delta\Delta G^{\text{true}}_{ij,\ \text{target}} \tag{21}$$

From these ϵ values, we calculated the correlation coefficient, ρ, from the sampled errors for the finite set of molecules for which measurements were available,

$$\rho = \frac{\text{cov}(\epsilon_{\text{target1}}, \epsilon_{\text{target2}})}{\sigma_{\epsilon\ \text{target 1}}\ \sigma_{\epsilon\ \text{target 2}}} \tag{22}$$

where $\sigma_{\epsilon\ \text{target 2}}$ is the standard deviation of $\epsilon_{\text{ij,target}}$.

To quantify ρ from these calculations, the default NUTS sampler with `jitter+adapt_diag` initialization, 3 000 tuning steps, and the default target accept probability was used to draw 20 000 samples from the model.

644 ## Calculating the marginal distribution of speedup

645 To quantify the expected speedup from the calculations we ran, we utilized 10^4 replicates of the scheme
646 detailed above to calculate the speedup given parameters ρ, $\sigma_{\text{sys,ij,1}}$, and $\sigma_{\text{sys,ij,2}}$, in the regime of infinite
647 effort and zero statistical error. Using Numpy 1.14.2, ρ was drawn from a normal distribution with the
648 mean and standard deviation from the posterior distribution of ρ from the Bayesian Graphical model. The
649 per-target systematic errors, $\sigma_{\text{sys,ij,1}}$ and $\sigma_{\text{sys,ij,2}}$, were estimated from the mean of the absolute value of $\epsilon_{\text{ij,1}}$
650 and $\epsilon_{\text{ij,2}}$, which are the magnitude of errors from the Bayesian graphical model. $\sigma_{\text{selectivity}}$ was calculated
651 using Equation 3. 10^6 molecules were proposed from true normal distribution, as above. The error of the
652 computational method was modeled as in Equation 5.

Data Availability

All curated starting structures, FEP+ results, and data analysis scripts and notebooks are available on GitHub: https://github.com/choderalab/selectivity

Acknowledgments

The authors are grateful to Patrick Grinaway (ORCID: 0000-0002-9762-4201) for useful discussions about Bayesian statistics and Mehtap Işık (ORCID: 0000-0002-6789-952X) for useful discussion about kinase inhibitor protonation states. SKA is grateful to Haoyu S. Yu, Wei Chen, and Dmitry Lupyan for advice on running FEP+ calculations.

Funding

Research reported in this publication was supported by the National Institute for General Medical Sciences of the National Institutes of Health under award numbers R01GM121505 and P30CA008748. SKA acknowledges financial support from Schrödinger and the Sloan Kettering Institute. JDC acknowledges financial support from Cycle for Survival and the Sloan Kettering Institute.

Disclosures

JDC was a member of the Scientific Advisory Board for Schrödinger, LLC during part of this study. JDC is a current member of the Scientific Advisory Board of OpenEye Scientific Software and a consultant for Foresite Labs. The Chodera laboratory receives or has received funding from multiple sources, including the National Institutes of Health, the National Science Foundation, the Parker Institute for Cancer Immunotherapy, Relay Therapeutics, Bayer, Entasis Therapeutics, Silicon Therapeutics, EMD Serono (Merck KGaA), AstraZeneca, XtalPi, the Molecular Sciences Software Institute, the Starr Cancer Consortium, the Open systematic Consortium, Cycle for Survival, a Louis V. Gerstner Young Investigator Award, and the Sloan Kettering Institute. A complete funding history for the Chodera lab can be found at http://choderalab.org/funding

Author Contributions

Conceptualization: SKA, LW, RA, JDC; Methodology: SKA, LW, JDC; Formal Analysis: SKA, JDC, LW; Data Curation: SKA, SP; Investigation: SKA, SP, AV; Writing – Original Draft: SKA, JDC; Writing – Review & Editing: SKA, JDC, LW, AV, RA; Visualization: SKA, JDC, LW; Supervision: LW, JDC, RA; Project Administration: SKA, LW, JDC, RA; Funding Acquisition: RA, JDC; Resources: LW, JDC

References

[1] **Chodera JD**, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS. Alchemical free energy methods for drug discovery: progress and challenges. Curr Opin Struct Biol. 2011 Apr; 21(2):150–160.

[2] **Huang J**, MacKerell AD. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. J Comput Chem. 2013 Sep; 34(25):2135–2145. doi: 10.1002/jcc.23354.

[3] **Maier JA**, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015 Aug; 11(8):3696–3713. doi: 10.1021/acs.jctc.5b00255.

[4] **Harder E**, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL, Kaus JW, Cerutti DS, Krilov G, Jorgensen WL, Abel R, Friesner RA. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. J Chem Theory Comput. 2016 Jan; 12(1):281–296.

[5] **Cournia Z**, Allen B, Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. Journal of chemical information and modeling. 2017 Dec; 57(12):2911–2937.

[6] **Brown SP**, Muchmore SW, Hajduk PJ. Healthy Skepticism: Assessing Realistic Model Performance. Drug Discov Today. 2009; 14(7):420 – 427. doi: http://dx.doi.org/10.1016/j.drudis.2009.01.012.

[7] **Abel R**, Mondal S, Masse C, Greenwood J, Harriman G, Ashwell MA, Bhat S, Wester R, Frye L, Kapeller R, Friesner RA. Accelerating drug discovery through tight integration of expert molecular design and predictive scoring. Curr Opin Struct Biol. 2017 Apr; 43(Supplement C):38–44.

[8] **Lovering F**, Aevazelis C, Chang J, Dehnhardt C, Fitz L, Han S, Janz K, Lee J, Kaila N, McDonald J, Moore W, Moretto A, Papaioannou N, Richard D, Ryan MS, Wan ZK, Thorarensen A. Imidazotriazines: Spleen Tyrosine Kinase (Syk) Inhibitors Identified by Free-Energy Perturbation (FEP). ChemMedChem. 2016 Jan; 11(2):217–233.

[9] **Ciordia M**, Pérez-Benito L, Delgado F, Trabanco AA, Tresadern G. Application of Free Energy Perturbation for the Design of BACE1 Inhibitors. Journal of chemical information and modeling. 2016 Sep; 56(9):1856–1871.

[10] **Lenselink EB**, Louvel J, Forti AF, van Veldhoven JPD, de Vries H, Mulder-Krieger T, McRobb FM, Negri A, Goose J, Abel R, van Vlijmen HWT, Wang L, Harder E, Sherman W, IJzerman AP, Beuming T. Predicting Binding Affinities for GPCR Ligands Using Free-Energy Perturbation. ACS omega. 2016 Aug; 1(2):293–304.

[11] **Jorgensen WL**. Computer-aided discovery of anti-HIV agents. Bioorganic & medicinal chemistry. 2016 Oct; 24(20):4768–4778.

[12] **Wang L**, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. J Am Chem Soc. 2015 Feb; 137(7):2695–2703. doi: 10.1021/ja512751q.

[13] **Abel R**, Wang L, Harder ED, Berne BJ, Friesner RA. Advancing Drug Discovery through Enhanced Free Energy Calculations. Accounts of chemical research. 2017 Jul; 50(7):1625–1632.

[14] **Zhang J**, Yang PL, Gray NS. Targeting cancer with small molecule kinase inhibitors. Nat Rev Cancer. 2009 Jan; 9(1):28–39.

[15] **Huggins DJ**, Sherman W, Tidor B. Rational approaches to improving selectivity in drug design. J Med Chem. 2012 Feb; 55(4):1424–1444.

[16] **Fan QW**, Cheng CK, Nicolaides TP, Hackett CS, Knight ZA, Shokat KM, Weiss WA. A dual phosphoinositide-3-kinase alpha/mTOR inhibitor cooperates with blockade of epidermal growth factor receptor in PTEN-mutant glioma. Cancer Res. 2007 Sep; 67(17):7960–7965.

[17] **Apsel B**, Blair JA, Gonzalez B, Nazif TM, Feldman ME, Aizenstein B, Hoffman R, Williams RL, Shokat KM, Knight ZA. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. Nat Chem Biol. 2008 Nov; 4(11):691–699.

[18] **Knight ZA**, Lin H, Shokat KM. Targeting the Cancer Kinome through Polypharmacology. Nat Rev Cancer. 2010; 10(2):130.

[19] **Hopkins AL**, Mason JS, Overington JP. Can we rationally design promiscuous drugs? Curr Opin Struct Biol. 2006 Feb; 16(1):127–136.

[20] **Hopkins AL**. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol. 2008 Nov; 4(11):682–690.

[21] **Kijima T**, Shimizu T, Nonen S, Furukawa M, Otani Y, Minami T, Takahashi R, Hirata H, Nagatomo I, Takeda Y, Kida H, Goya S, Fujio Y, Azuma J, Tachibana I, Kawase I. Safe and successful treatment with erlotinib after gefitinib-induced hepatotoxicity: difference in metabolism as a possible mechanism. J Clin Oncol. 2011 Jul; 29(19):e588–90.

[22] **Liu S**, Kurzrock R. Toxicity of targeted therapy: Implications for response and impact of genetic polymorphisms. Cancer Treat Rev. 2014 Aug; 40(7):883–891.

[23] **Rudmann DG**. On-target and off-target-based toxicologic effects. Toxicol Pathol. 2013 Feb; 41(2):310–314.

[24] **Mendoza MC**, Er EE, Blenis J. The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. Trends Biochem Sci. 2011 Jun; 36(6):320–328.

[25] **Tricker EM**, Xu C, Uddin S, Capelletti M, Ercan D, Ogino A, Pratilas CA, Rosen N, Gray NS, Wong KK, Jänne PA. Combined EGFR/MEK Inhibition Prevents the Emergence of Resistance in EGFR-Mutant Lung Cancer. Cancer Discov. 2015 Sep; 5(9):960–971.

[26] **Bailey ST**, Zhou B, Damrauer JS, Krishnan B, Wilson HL, Smith AM, Li M, Yeh JJ, Kim WY. mTOR Inhibition Induces Compensatory, Therapeutically Targetable MEK Activation in Renal Cell Carcinoma. PLoS One. 2014 Sep; 9(9):e104413.

[27] **Chandarlapaty S**, Sawai A, Scaltriti M, Rodrik-Outmezguine V, Grbovic-Huezo O, Serra V, Majumder PK, Baselga J, Rosen N. AKT Inhibition Relieves Feedback Suppression of Receptor Tyrosine Kinase Expression and Activity. Cancer Cell. 2011 Jan; 19(1):58–71. doi: 10.1016/j.ccr.2010.10.031.

[28] **Pao W**, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, Singh B, Heelan R, Rusch V, Fulton L, Mardis E, Kupfer D, Wilson R, Kris M, Varmus H. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proceedings of the National Academy of Sciences. 2004 Sep; 101(36):13306–13311.

[29] **Kim Y**, Li Z, Apetri M, Luo B, Settleman JE, Anderson KS. Temporal resolution of autophosphorylation for normal and oncogenic forms of EGFR and differential effects of gefitinib. Biochemistry. 2012 Jun; 51(25):5212–5222.

[30] **Juchum M**, Günther M, Laufer SA. Fighting Cancer Drug Resistance: Opportunities and Challenges for Mutation-Specific EGFR Inhibitors. Drug Resist Updat. 2015 May; 20:12–28. doi: 10.1016/j.drup.2015.05.002.

[31] **Din OS**, Woll PJ. Treatment of gastrointestinal stromal tumor: focus on imatinib mesylate. Ther Clin Risk Manag. 2008 Feb; 4(1):149–162.

[32] **Lin YL**, Meng Y, Jiang W, Roux B. Explaining why Gleevec is a specific and potent inhibitor of Abl kinase. Proc Natl Acad Sci U S A. 2013 Jan; 110(5):1664–1669.

[33] **Lin YL**, Meng Y, Huang L, Roux B. Computational Study of Gleevec and G6G Reveals Molecular Determinants of Kinase Inhibitor Selectivity. J Am Chem Soc. 2014 Oct; 136(42):14753–14762.

[34] **Lin YL**, Roux B. Computational Analysis of the Binding Specificity of Gleevec to Abl, c-Kit, Lck, and c-Src Tyrosine Kinases. J Am Chem Soc. 2013 Oct; 135(39):14741–14753.

[35] **Aldeghi M**, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. J Am Chem Soc. 2017 Jan; 139(2):946–957.

[36] **Moraca F**, Negri A, de Oliveira C, Abel R. Application of Free Energy Perturbation (FEP+) to Understanding Ligand Selectivity: A Case Study to Assess Selectivity Between Pairs of Phosphodiesterases (PDE's). Journal of Chemical Information and Modeling. 2019; 59(6):2729–2740. https://doi.org/10.1021/acs.jcim.9b00106, doi: 10.1021/acs.jcim.9b00106, pMID: 31144815.

[37] **Robert Roskoski Jr**. USFDA Approved Protein Kinase Inhibitors. . 2017; http://www.brimr.org/PKI/PKIs.htm, updated 3 May 2017.

[38] **Santos R**, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, Karlsson A, Al-Lazikani B, Hersey A, Oprea TI, Overington JP. A Comprehensive Map of Molecular Drug Targets. Nat Rev Drug Discov. 2016 Dec; 16(1):19–34. doi: 10.1038/nrd.2016.230.

[39] **Volkamer A**, Eid S, Turk S, Jaeger S, Rippmann F, Fulle S. Pocketome of human kinases: prioritizing the ATP binding sites of (yet) untapped protein kinases for drug discovery. J Chem Inf Model. 2015 Mar; 55(3):538–549.

[40] **Manning G**, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The Protein Kinase Complement of the Human Genome. Science. 2002 Dec; 298(5600):1912–1934.

[41] **Wu P**, Nielsen TE, Clausen MH. FDA-approved small-molecule kinase inhibitors. Trends Pharmacol Sci. 2015 Jul; 36(7):422–439.

[42] **Cowan-Jacob SW**, Fendrich G, Floersheimer A, Furet P, Liebetanz J, Rummel G, Rheinberger P, Centeleghe M, Fabbro D, Manley PW, IUCr. Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. Acta Crystallogr D Biol Crystallogr. 2007 Jan; 63(1):80–93.

[43] **Seeliger MA**, Nagar B, Frank F, Cao X, Henderson MN, Kuriyan J. c-Src Binds to the Cancer Drug Imatinib with an Inactive Abl/c-Kit Conformation and a Distributed Thermodynamic Penalty. Structure. 2007 Mar; 15(3):299–311.

[44] **Huse M**, Kuriyan J. The conformational plasticity of protein kinases. Cell. 2002 Jan; 109(3):275–282.

[45] **Harrison SC**. Variation on an Src-like theme. Cell. 2003 Mar; 112(6):737–740.

[46] **Volkamer A**, Eid S, Turk S, Rippmann F, Fulle S. Identification and Visualization of Kinase-Specific Subpockets. J Chem Inf Model. 2016 Feb; 56(2):335–346.

[47] **Christmann-Franck S**, van Westen GJP, Papadatos G, Beltran Escudie F, Roberts A, Overington JP, Domine D. Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design? Journal of chemical information and modeling. 2016 Sep; 56(9):1654–1675.

[48] **Anastassiadis T**, Deacon SW, Devarajan K, Ma H, Peterson JR. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat Biotechnol. 2011 Nov; 29(11):1039–1045.

[49] **Davis MI**, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP. Comprehensive Analysis of Kinase Inhibitor Selectivity. Nat Biotechnol. 2011 Oct; 29(11):1046–1051. doi: 10.1038/nbt.1990.

[50] **Klaeger S**, Heinzlmeir S, Wilhelm M, Polzer H, Vick B, Koenig PA, Reinecke M, Ruprecht B, Petzoldt S, Meng C, Zecha J, Reiter K, Qiao H, Helm D, Koch H, Schoof M, Canevari G, Casale E, Depaolini SR, Feuchtinger A, et al. The target landscape of clinical kinase drugs. Science. 2017 Dec; 358(6367).

[51] **Sun C**, Hobor S, Bertotti A, Zecchin D, Huang S, Galimi F, Cottino F, Prahallad A, Grernrum W, Tzani A, Schlicker A, Wessels LFA, Smit EF, Thunnissen E, Halonen P, Lieftink C, Beijersbergen RL, Di Nicolantonio F, Bardelli A, Trusolino L, et al. Intrinsic resistance to MEK inhibition in KRAS mutant lung and colon cancer through transcriptional induction of ERBB3. CellReports. 2014 Apr; 7(1):86–93.

[52] **Manchado E**, Weissmueller S, Morris JP, Chen CC, Wullenkord R, Lujambio A, de Stanchina E, Poirier JT, Gainor JF, Corcoran RB, Engelman JA, Rudin CM, Rosen N, Lowe SW. A combinatorial strategy for treating KRAS-mutant lung cancer. Nature. 2016 Jun; 534(7609):647–651.

[53] **Shao H**, Shi S, Huang S, Hole AJ, Abbas AY, Baumli S, Liu X, Lam F, Foley DW, Fischer PM, Noble M, Endicott JA, Pepper C, Wang S. Substituted 4-(Thiazol-5-yl)-2-(phenylamino)pyrimidines Are Highly Active CDK9 Inhibitors: Synthesis, X-ray Crystal Structures, Structure–Activity Relationship, and Anticancer Activities. J Med Chem. 2013 Feb; 56(3):640–659.

[54] **Blake JF**, Burkard M, Chan J, Chen H, Chou KJ, Diaz D, Dudley DA, Gaudino JJ, Gould SE, Grina J, Hunsaker T, Liu L, Martinson M, Moreno D, Mueller L, Orr C, Pacheco P, Qin A, Rasor K, Ren L, et al. Discovery of (S)-1-(1-(4-Chloro-3-fluorophenyl)-2-hydroxyethyl)-4-(2-((1-methyl-1H-pyrazol-5-yl)amino)pyrimidin-4-yl)pyridin-2(1H)-one (GDC-0994), an Extracellular Signal-Regulated Kinase 1/2 (ERK1/2) Inhibitor in Early Clinical Development. J Med Chem. 2016 Jun; 59(12):5650–5660.

[55] **Bosc N**, Meyer C, Bonnet P. The use of novel selectivity metrics in kinase research. BMC bioinformatics. 2017 Jan; 18(1):17.

[56] **Cheng AC**, Eksterowicz J, Geuns-Meyer S, Sun Y. Analysis of kinase inhibitor selectivity using a thermodynamics-based partition index. J Med Chem. 2010 Jun; 53(11):4502–4510.

[57] **Shirts MR**, Mobley DL, Brown SP. Free energy calculations in structure-based drug design. In: *Structure Based Drug Design* Cambridge University Press; 2009.

[58] **Kooistra AJ**, Kanev GK, van Linden OPJ, Leurs R, de Esch IJP, de Graaf C. KLIFS: a structural kinase-ligand interaction database. Nucleic acids research. 2016 Jan; 44(D1):D365–D371.

[59] **van Linden OPJ**, Kooistra AJ, Leurs R, de Esch IJP, de Graaf C. KLIFS: a knowledge-based structural database to navigate kinase-ligand interaction space. Journal of medicinal chemistry. 2014 Jan; 57(2):249–277.

[60] **Hole AJ**, Baumli S, Shao H, Shi S, Huang S, Pepper C, Fischer PM, Wang S, Endicott JA, Noble ME. Comparative Structural and Functional Studies of 4-(Thiazol-5-yl)-2-(phenylamino)pyrimidine-5-carbonitrile CDK9 Inhibitors Suggest the Basis for Isotype Selectivity. J Med Chem. 2013 Feb; 56(3):660–670.

[61] **Yung-Chi C**, Prusoff WH. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochemical Pharmacology. 1973; 22(23):3099 – 3108. http://www.sciencedirect.com/science/article/pii/0006295273901962, doi: https://doi.org/10.1016/0006-2952(73)90196-2.

[62] **Hauser K**, Negron C, Albanese SK, Ray S, Steinbrecher T, Abel R, Chodera JD, Wang L. Predicting resistance of clinical Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. Communications Biology. 2018 Jun; 1(1):70.

[63] **Kalliokoski T**, Kramer C, Vulpetti A, Gedeck P. Comparability of Mixed IC50 Data–a Statistical Analysis. PloS One. 2013; 8(4):e61007.

[64] **Hari SB**, Merritt EA, Maly DJ. Sequence determinants of a specific inactive protein kinase conformation. Chemistry & biology. 2013 Jun; 20(6):806–815.

[65] **Davis MI**, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol. 2011 Oct; 29(11):1046–1051.

[66] **Graczyk PP**. Gini coefficient: a new way to express selectivity of kinase inhibitors against a family of kinases. Journal of medicinal chemistry. 2007 Nov; 50(23):5773–5779.

[67] **Duong-Ly KC**, Devarajan K, Liang S, Horiuchi KY, Wang Y, Ma H, Peterson JR. Kinase Inhibitor Profiling Reveals Unexpected Opportunities to Inhibit Disease-Associated Mutant Kinases. CellReports. 2016 Feb; 14(4):772–781.

[68] **Uitdehaag JCM**, Zaman GJR. A theoretical entropy score as a single value to express inhibitor selectivity. BMC bioinformatics. 2011 Apr; 12:94.

[69] **Michel J**, Verdonk ML, Essex JW. Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization? Journal of medicinal chemistry. 2006 Dec; 49(25):7427–7439.

[70] **Berman HM**, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr. 2002 Jun; 58(Pt 61):899–907.

[71] **Sastry GM**, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des. 2013 Mar; 27(3):221–234.

[72] **Salvatier J**, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Computer Science. 2016; 2:e55.

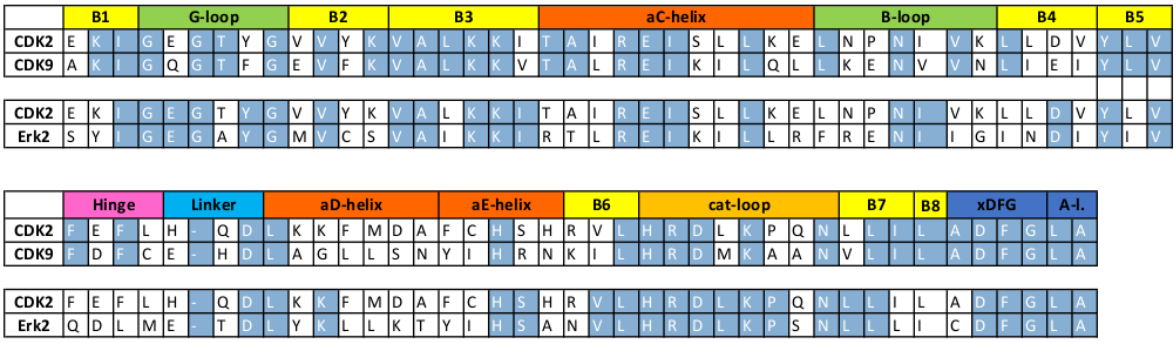
[73] **Al-Rfou R**, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F, Bayer J, Belikov A, Belopolsky A, Bengio Y, Bergeron A, Bergstra J, Bisson V, Bleecher Snyder J, Bouchard N, Boulanger-Lewandowski N, Bouthillier X, de Brébisson A, Breuleux O, et al. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints. 2016 May; abs/1605.02688. http://arxiv.org/abs/1605.02688.

[74] **Hu J**, Ahuja LG, Meharena HS, Kannan N, Kornev AP, Taylor SS, Shaw AS. Kinase regulation by hydrophobic spine assembly in cancer. Molecular and cellular biology. 2015 Jan; 35(1):264–276.

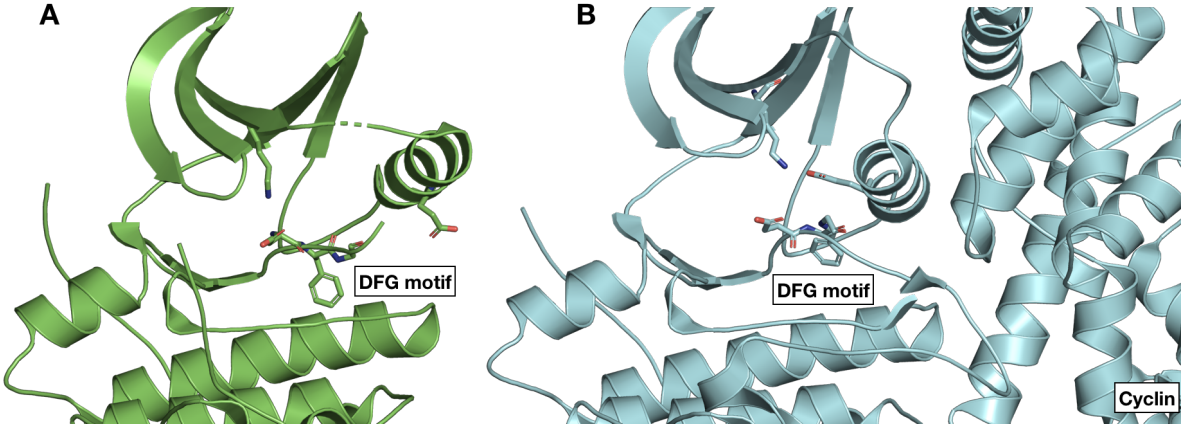
Supplemental Information

Supplemental Table. 1. The CDK2 and CDK9 binding sites are more similar than the CDK2 and ERK2 binding sites
Sequence based similarity of the binding sites based on multiple sequence alignments of the 85 residues annotated by the KLIFS. Upper triangle shows the ratio of sequence identity, lower triangle, the number of matching residues out of 85 binding site residues. database [58, 59]

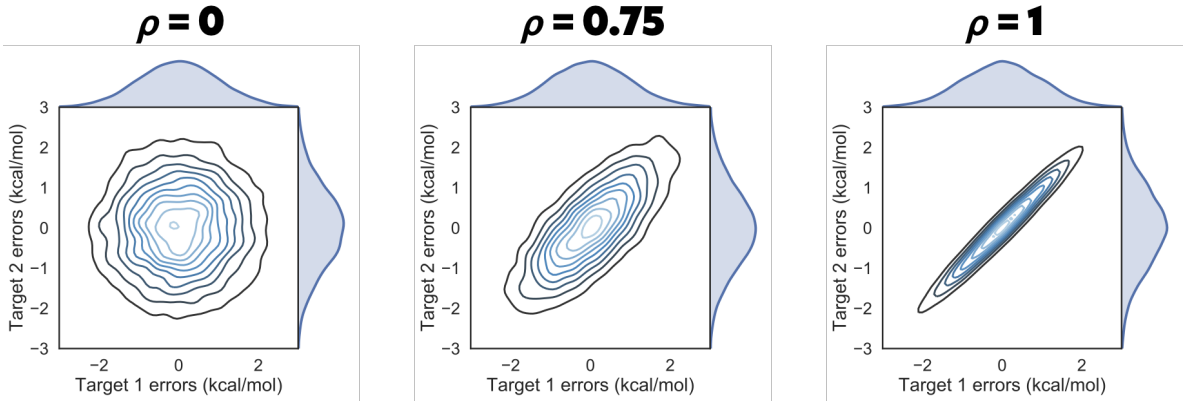
Kinase	CDK2	CDK9	ERK2
CDK2	1.0	0.57	0.52
CDK9	47	1.0	0.52
ERK2	43	43	1.0



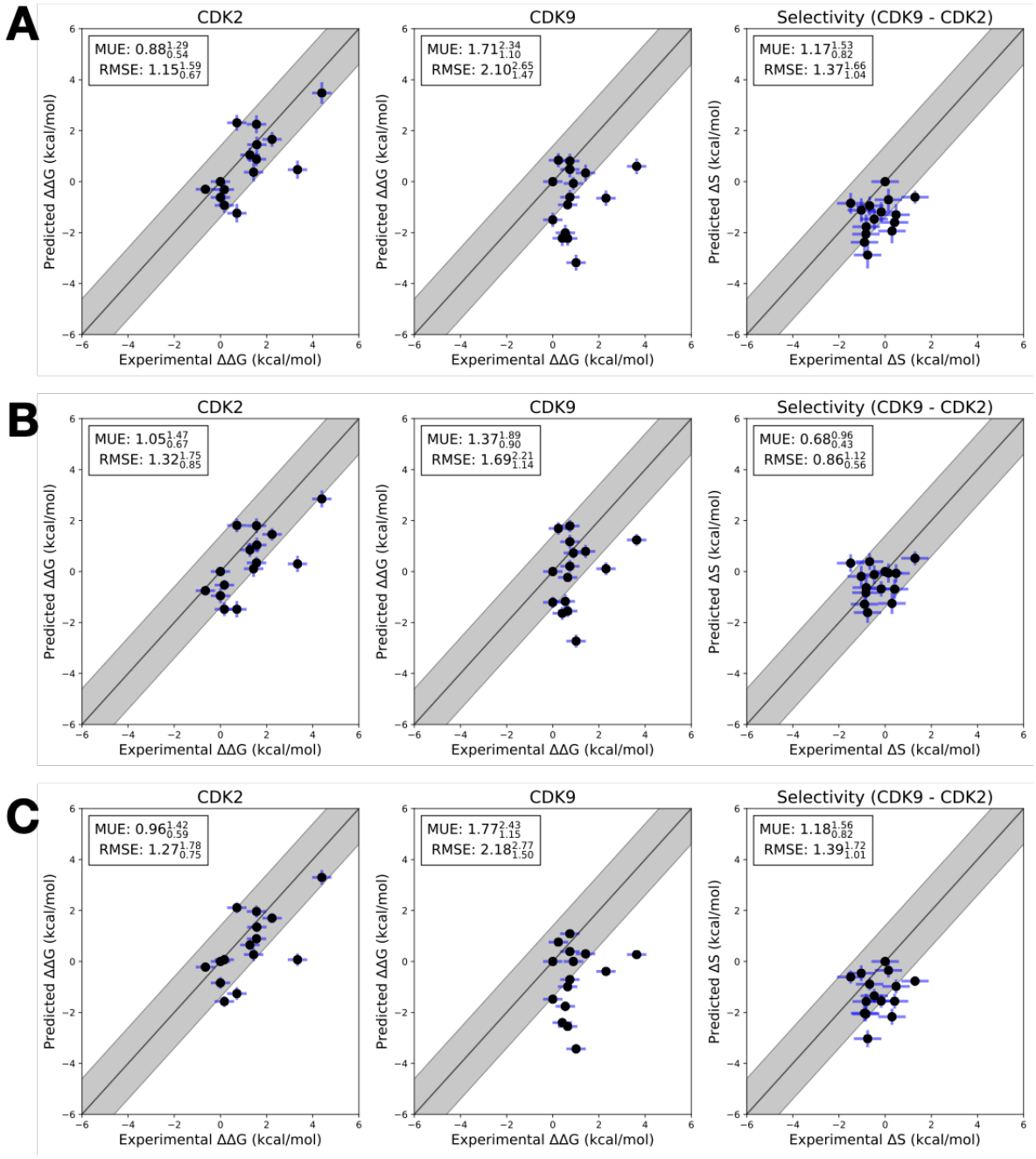
Supplemental Figure. 1. Sequence identity of the kinase pairs by binding site motif
Binding site sequence identity for CDK2/CDK9 and CDK2/ERK2 by binding site motif, as defined by the KLIFs database [58, 59]. Identical residues between the pairs are shown in blue.



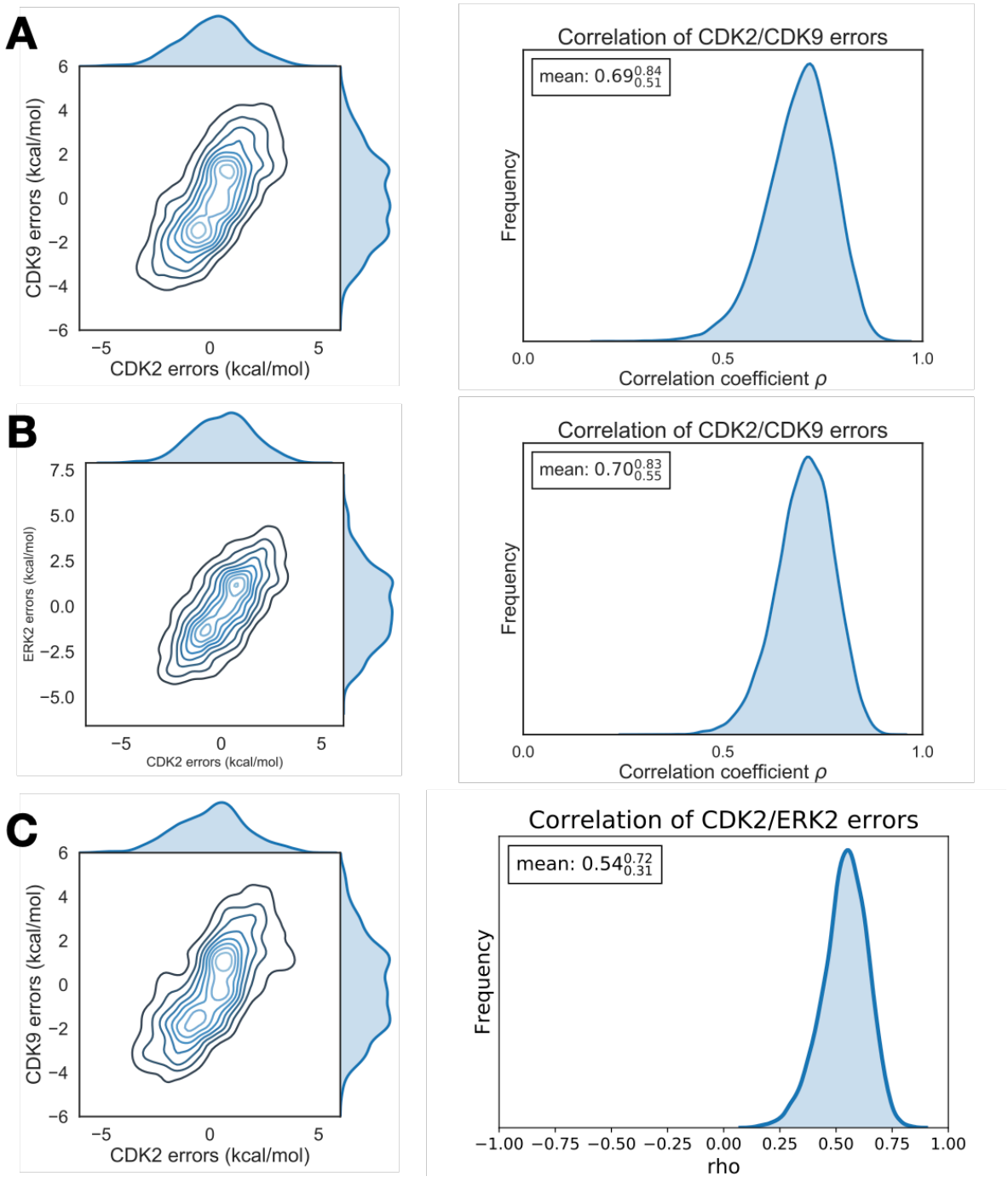
Supplemental Figure. 2. CDK2 adopts an inactive conformation in the crystal structure used for the CDK2/ERK2 calculations
(**A**) CDK2 (5K4J) adopts an inactive conformation in the absence of its cyclin. The DFG motif is in a DFG-in conformation, with the αC helix rotated outwards, breaking the salt bridge between K33 and E51 (Uniprot numbering) that is typically a marker of an active conformation. Notably, the Phe in the DFG motif does not completely form the hydrophobic spine due to the rotation of the αC helix [74] (**B**) The CDK2 structure used for the CDK2/CDK9 calculations (4BCK) contains cyclin A and adopts a DFG-in/αC helix-in conformation that forms the salt bridge between K33 and E51. This is typically indicative of a fully active kinase [44, 64].



Supplemental Figure. 3. Correlation coefficient ρ controls the shape of the joint marginal distribution of errors
As ρ increases, the joint marginal distribution of errors become more diagonal. Each panel shows 10000 samples drawn from a multivariate normal distribution centered around 0 kcal/mol, where the per target error was set to 1 kcal/mol and ρ to the value indicated in bold over the plot.

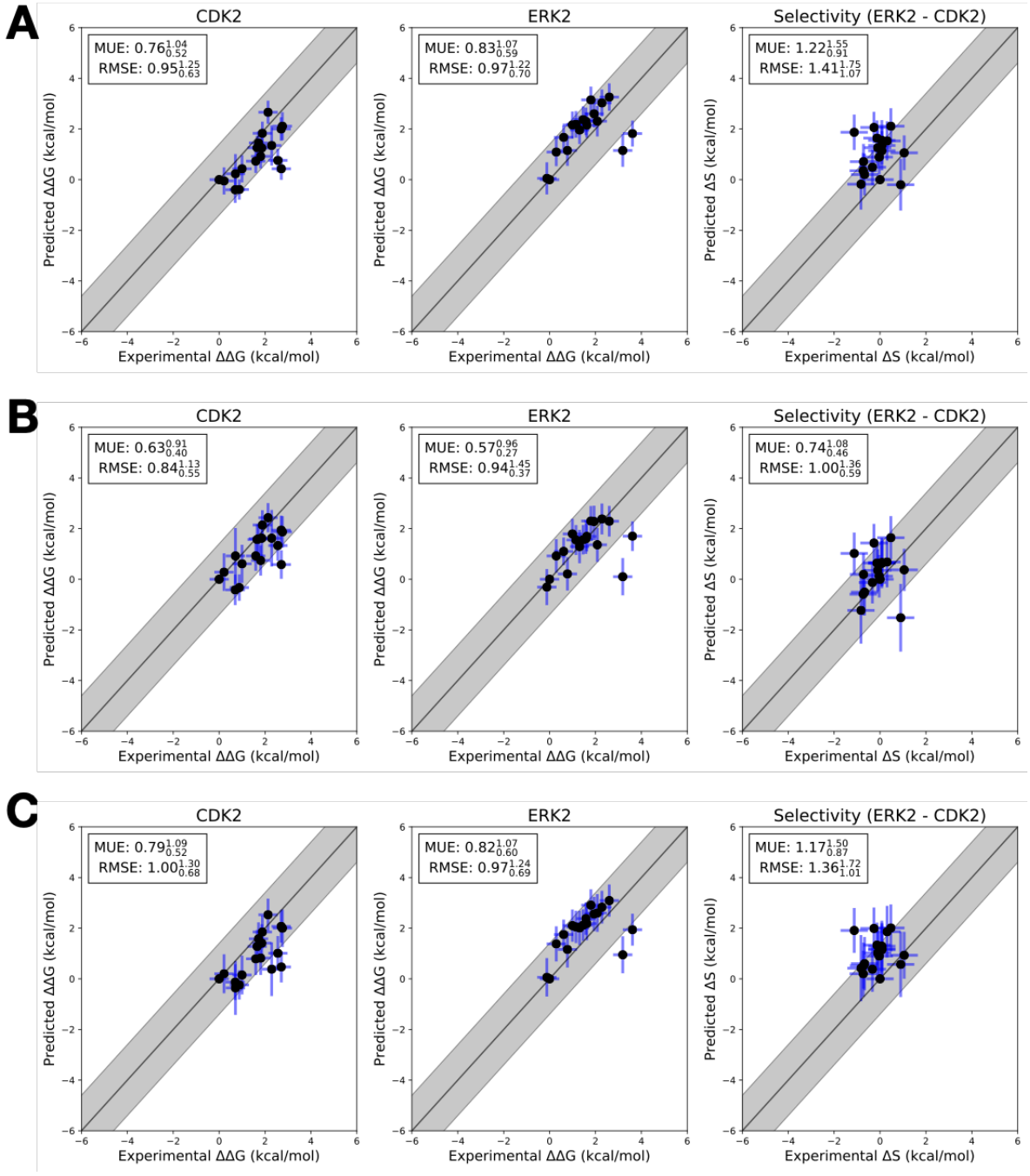


Supplemental Figure. 4. Each replicate of the CDK2/CDK9 calculations yields a consistent RMSE and MUE
Three replicates of the CDK2/CDK9 calculations with different random seeds, but otherwise the same input structures, files, and parameters. The experimental values are shown on the X-axis and calculated values on the Y-axis. Each data point corresponds to a transformation between a ligand i to a set reference ligand j (Compound 1a) for a given target. All values are shown in units of kcal/mol. The blue vertical error bars are $\sigma_{\text{stat,ij,target}}$, which was estimated by calculating the standard deviation of $\Delta\Delta G^{\text{FEP}}_{ij,\text{ target}}$ from the Bayesian model described in depth in **Methods**. The horizontal error bars show $\delta\Delta\Delta G^{exp}_{ij}$ based on the assumed uncertainty of 0.3 kcal/mol[6, 63] for each ΔG^{exp}_{i} expanded assuming no correlation between each measurement. For selectivity, the errors were propagated under the assumption that they were completely uncorrelated. The black line indicates agreement between calculation and experiment, while the gray shaded region represent 1.36 kcal/mol (or 1 log unit) error. The MUE and RMSE are shown on each plot with bootstrapped 95% confidence intervals. (**A**) Replicate 1 (**B**) Replicate 2 (**C**) Replicate 3

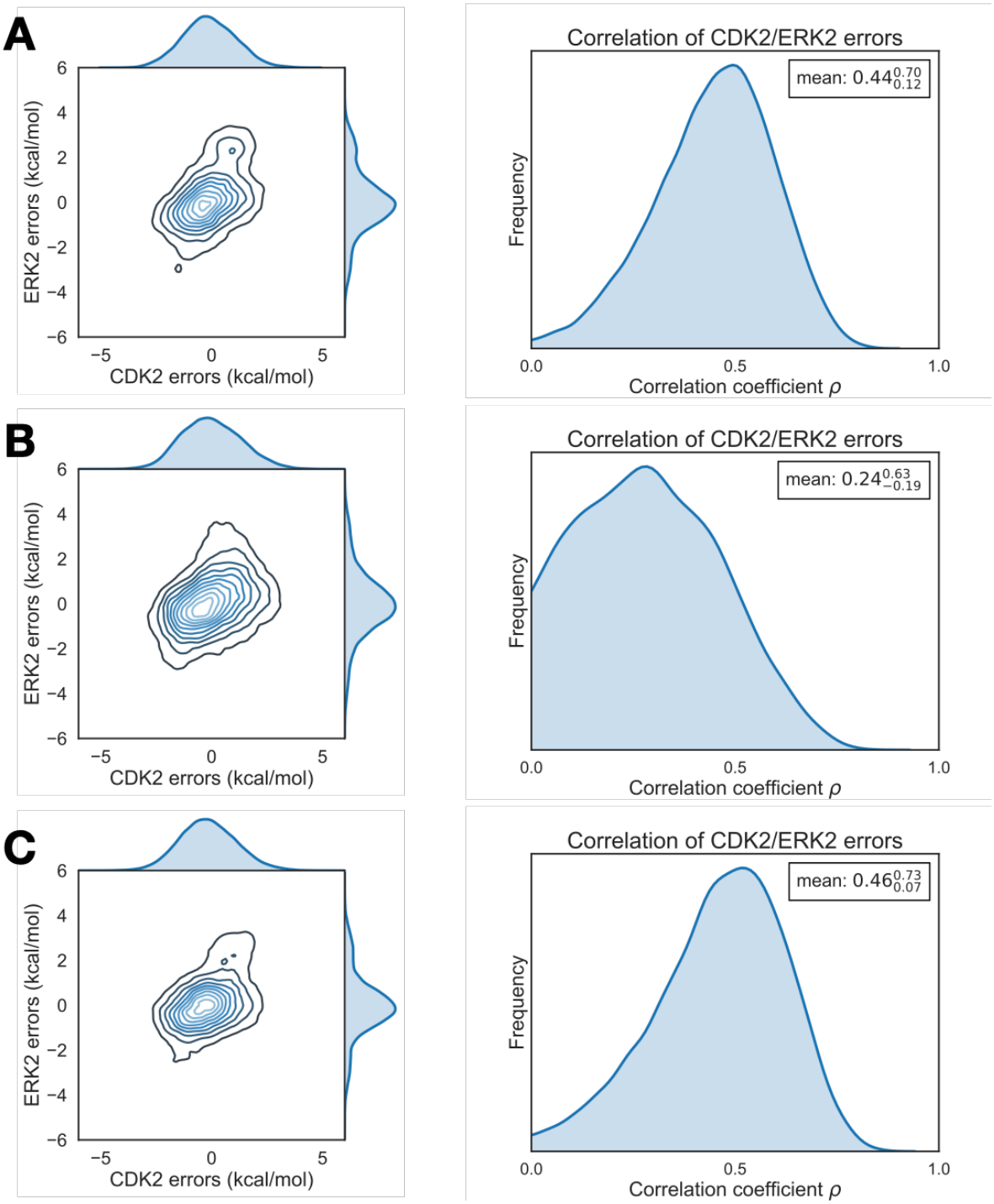


Supplemental Figure. 5. Each replicate of the CDK2/CDK9 calculations yields consistent errors and correlation coefficient

(**A**) (*left*) The joint posterior distribution of the prediction errors for CDK2 (X-axis) and CDK9 (Y-axis) from the Bayesian graphical model for replicate 1. (*right*) The posterior marginal distribution of the correlation coefficient (ρ) is shown in gray for replicate 1. The inserted box shows the mean and 95% confidence interval for the correlation coefficient. (**B**) and (**C**) The same as above, but for replicates 2 and 3, respectively

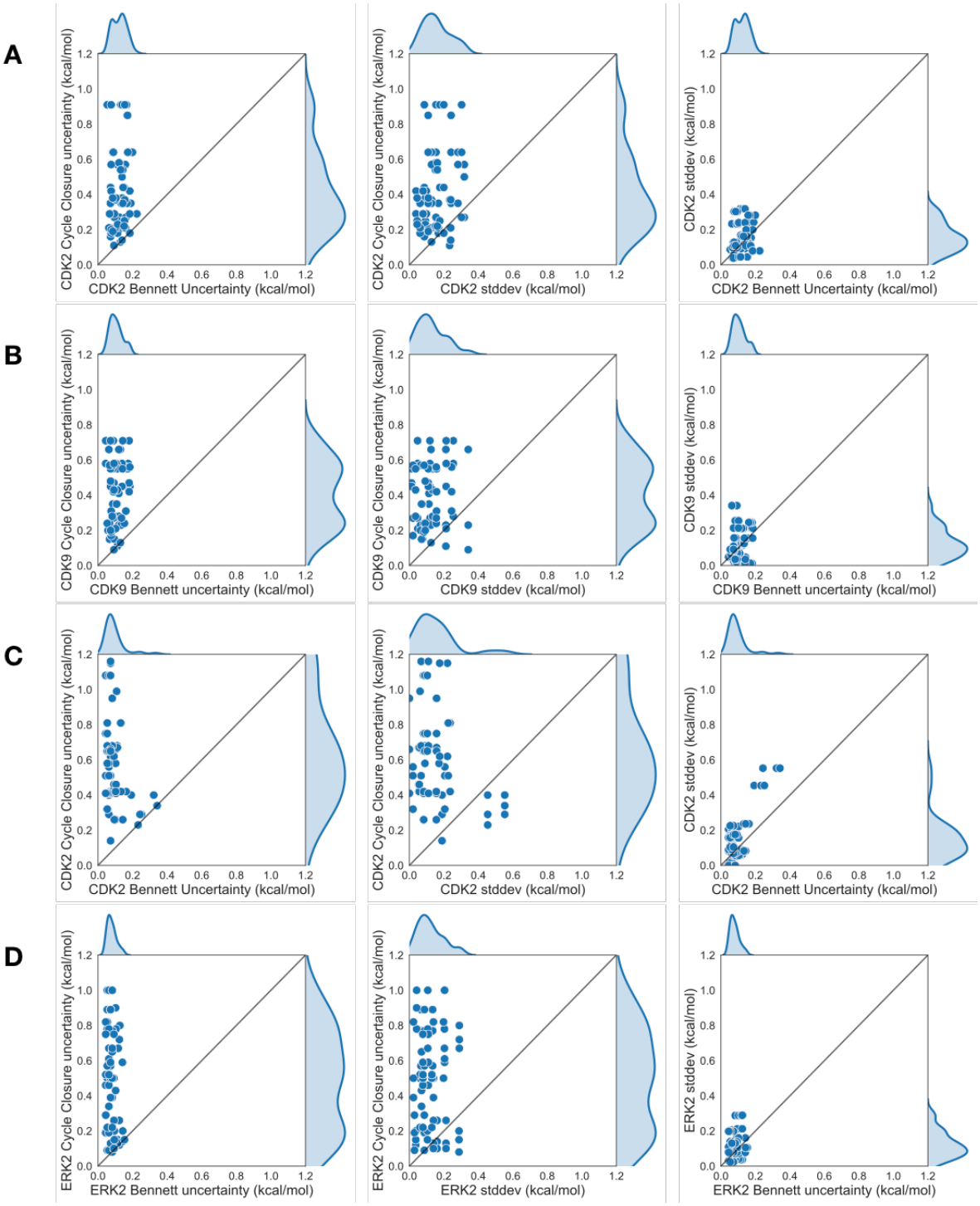


Supplemental Figure. 6. Each replicate of the CDK2/ERK2 calculations yields a consistent RMSE and MUE
Three replicates of the CDK2/ERK2 calculations with different random seeds, but otherwise the same input structures, files, and parameters. The experimental values are shown on the X-axis and calculated values on the Y-axis. Each data point corresponds to a transformation between a ligand i to reference ligand j (Compound 6) for a given target. All values are shown in units of kcal/mol. The blue vertical error bars are $\sigma_{\text{stat,ij,target}}$, which was estimated by calculating the standard deviation of $\Delta\Delta G^{\text{FEP}}_{ij,\text{ target}}$ from the Bayesian model described in depth in **Methods**. The horizontal error bars show $\delta\Delta\Delta G^{exp}_{ij}$ based on the assumed uncertainty of 0.3 kcal/mol[6, 63] for each ΔG^{exp}_{i} expanded assuming no correlation between each measurement. For selectivity, the errors were propagated under the assumption that they were completely uncorrelated. The black line indicates agreement between calculation and experiment, while the gray shaded region represent 1.36 kcal/mol (or 1 log unit) error. The MUE and RMSE are shown on each plot with bootstrapped 95% confidence intervals. (**A**) Replicate 1 (**B**) Replicate 2 (**C**) Replicate 3



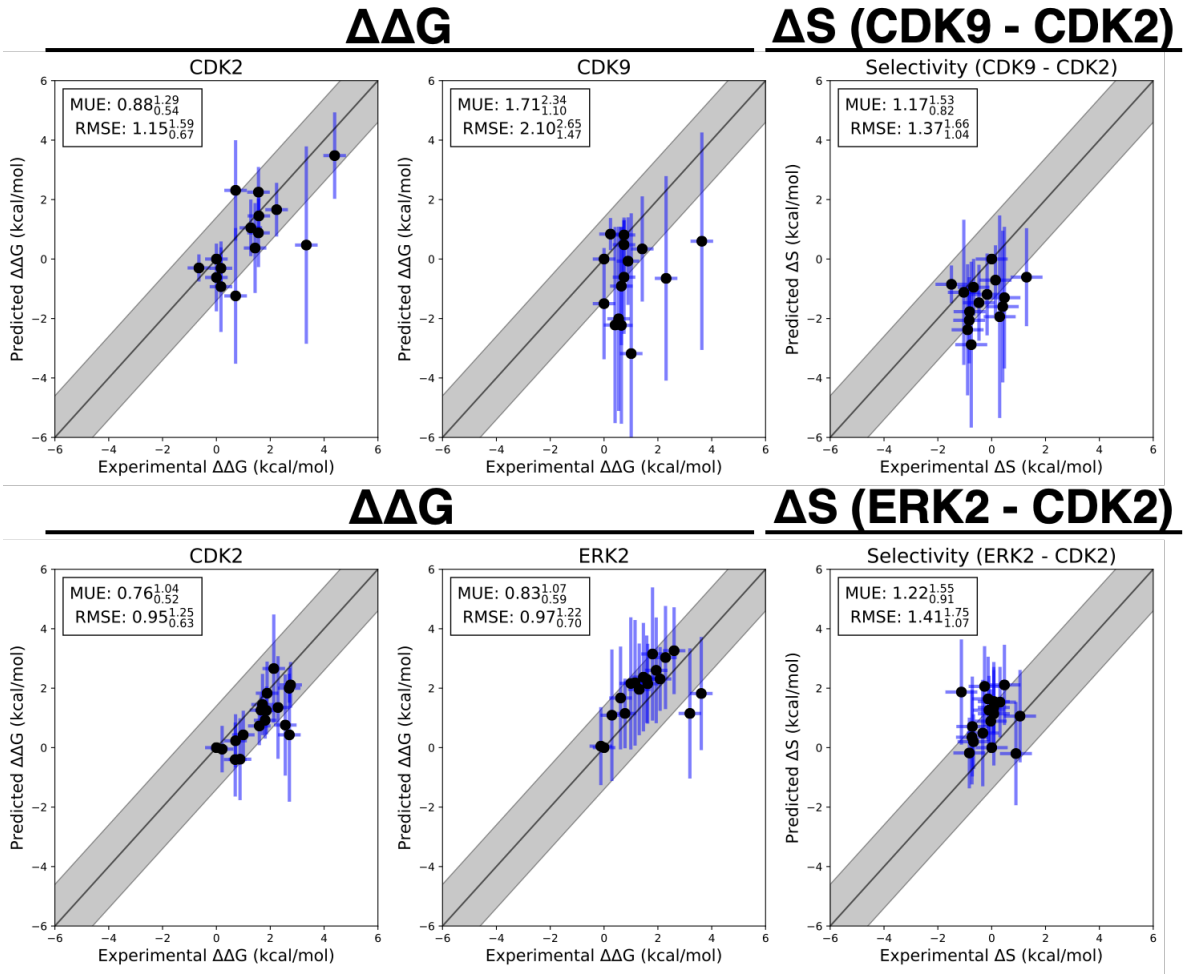
Supplemental Figure. 7. Each replicate of the CDK2/ERK2 calculations yields consistent errors and correlation coefficient

(**A**) (*left*) The joint posterior distribution of the prediction errors for CDK2 (X-axis) and ERK2 (Y-axis) from the Bayesian graphical model for replicate 1. (*right*) The posterior marginal distribution of the correlation coefficient (ρ) is shown in gray for replcicate 1. The inserted box shows the mean and 95% confidence interval for the correlation coefficient. (**B**) and (**C**) The same as above, but for replicates 2 and 3, respectively

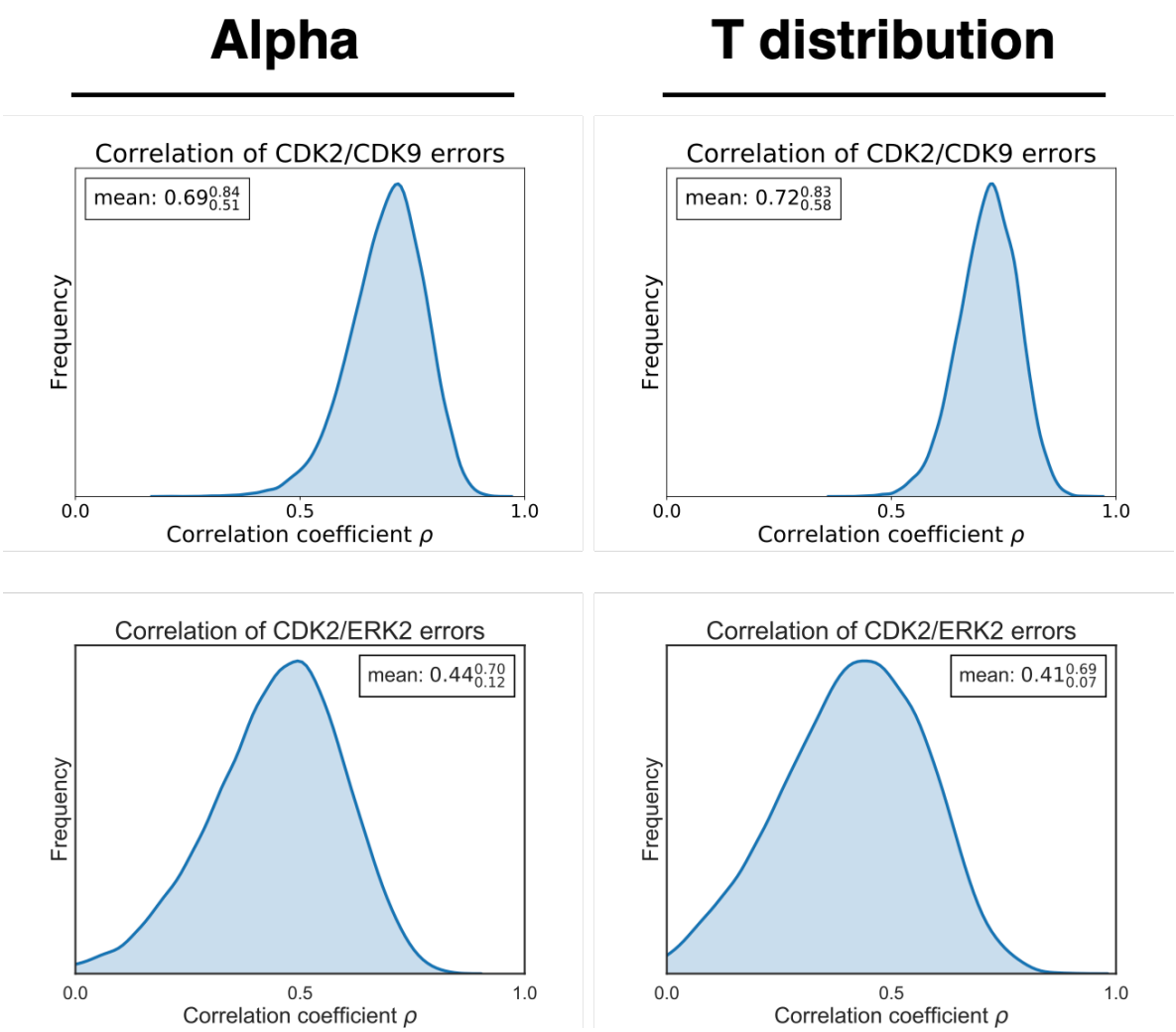


Supplemental Figure. 8. The standard deviation and Bennett error for each edge is smaller than the estimated cycle closure uncertainties

Pairwise Comparisons of the cycle closure uncertainty, the Bennett uncertainty, and the standard deviation of three replicate calculations, reported in kcal/mol. Each point corresponds to an edge of the FEP map. The edges for all three replicates are pooled and shown together. (**A and B**) CDK2/CDK9 calculations from the Shao data set (**C and D**) CDK2/ERK2 calculations from the Blake data set



Supplemental Figure. 9. The statistical and systematic error explain deviation from experimental measurements $\Delta\Delta G_{ij,\text{target}}$ and ΔS_{ij} predictions for CDK2/ERK2 from the Blake data sets (*top*), and CDK2/CDK9 (*bottom*) from the Shao data sets. The experimental values are shown on the X-axis and calculated values on the Y-axis. Each data point corresponds to a transformation between a ligand i to a set reference ligand j for a given target. All values are shown in units of kcal/mol. The horizontal error bars show the $\delta\Delta\Delta G_{ij}^{exp}$ based on the assumed uncertainty of 0.3 kcal/mol[6, 63] for each ΔG_i^{exp}. We show the estimated error ($\sigma_{\text{stat,ij,target}}$ + $\sigma_{\text{sys,ij,target}}$) as vertical blue error bars, which are one standard error. $\sigma_{\text{stat,ij,target}}$ was estimated by calculating the standard deviation of $\Delta\Delta G_{ij,\text{ target}}^{\text{FEP}}$ from the Bayesian model described in depth in **Methods**. $\sigma_{\text{sys,ij,target}}$ was estimated from the mean of $\epsilon_{ij,\text{target}}$ described in equation 21. For the ΔS panels, $\sigma_{\text{selectivity}}$ (vertical blue error bars) was calculated according to Equation 3 using the estimated correlation coefficients from Figure 5.The black line indicates agreement between calculation and experiment, while the gray shaded region represent 1.36 kcal/mol (or 1 log unit) error. The MUE and RMSE are shown on each plot with bootstrapped 95% confidence intervals.



Supplemental Figure. 10. Estimates of correlation coefficient ρ are insensitive to use of scaling term α or Student's t-distribution

(*left*) Estimate of correlation coefficient ρ for replicate 1 of the CDK2/CDK9 (*top*) and CDK2/ERK2 (*bottom*) calculations using a scaling term (α) to account for the BAR error underestimating the cycle closure statistical error, shown in greater detail in the **Methods** section Equation 17. (*right*) Estimate of correlation coefficient ρ for CDK2/CDK9 (*top*) and CDK2/ERK2 (*bottom*) using a Student's t-distribution instead of a Normal distribution and scaling term in Equation 17