

Is structure based drug design ready for selectivity optimization?

³ Steven K. Albanese^{1,2}, John D. Chodera², Simon Peng³, Robert Abel³, Lingle Wang^{3*}

⁴ ¹Louis V. Gerstner, Jr. Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer
⁵ Center, New York, NY 10065; ²Computational and Systems Biology Program, Sloan Kettering
⁶ Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065; ³Schrödinger, New York,
⁷ NY 10036

⁸ *For correspondence: lingle.wang@schrodinger.com (LW)

Abstract

Alchemical free energy calculations are now widely used to drive or maintain potency in small molecule lead optimization, where the binding affinity to a protein target can be computed—in well-behaved cases—to roughly 1 kcal/mol inaccuracy, which is believed to primarily stem from force field errors. Despite this, the potential to use free energy calculations to drive optimization of compound *selectivity* among two similar targets has been relatively unexplored. In the most optimistic scenario, the similarity of binding sites might lead to a fortuitous cancellation of force field 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 novel Bayesian analysis approach, we separate force field error from statistical error and quantify the correlation in force field errors between selectivity targets. We find that, in the closely related CDK2/CDK9 case, a high correlation in force field errors suggests free energy calculations can have significant impact in aiding chemists in achieving selectivity, while in more distantly related kinases (CDK2/ERK2), limited correlation in force field errors reduces the ability for free energy calculations to aid selectivity optimization. In both cases, the correlation in force field error suggests that longer simulations are beneficial to properly balance statistical error with systematic error to take full advantage of the increase in accuracy in selectivity prediction possible due to fortuitous cancellation of error.

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 prediction of ligand binding free energies, including all enthalpic and entropic contributions [1]. Advances in atomistic molecular mechanics forcefields and free energy methodologies [2–5] have allowed free energy methods to reach a level of accuracy sufficient for predicting ligand potencies [6]. Free energy 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 found that the use of FEP+ [12] to predict potency for 92 different projects and 3021 compounds found a median RMSE of 1 kcal/mol [13].

Selectivity is an important consideration in drug design

In addition to maintaining or optimizing potency, free energy methods can be applied to predicting the selectivity of a ligand between two or more targets. Selectivity is an important property to consider in drug development, either in the pursuit of a maximally selective inhibitor [14, 15] or in pursuit a polypharmacological agent [16–20], to avoid on-target toxicity (arising from inhibition of the intended target) [21] and off-target toxicity (arising from inhibition of unintended targets) [22, 23]. In either paradigm, considering the

43 selectivity of a compound is complicated by the biology of the target. For example, kinases exist as nodes
44 in complex signaling networks [24, 25] with feedback inhibition and cross-talk between pathways. Careful
45 consideration of which off-targets are being inhibited can avoid off-target toxicity due to alleviating feedback
46 inhibition and inadvertently reactivating the targeted pathway [24, 25], or the upregulation of a secondary
47 pathway by alleviation of cross-talk inhibition [26, 27]. Off-target toxicity can also be caused by inhibiting
48 unrelated targets, such as gefitinib, an EGFR inhibitor, inhibiting CYP2D6 [22] and causing hepatotoxicity
49 in lung cancer patients. In a cancer setting, on-target toxicity can be avoided by considering the selectivity
50 for the oncogenic mutant form of the kinase over the wild type form of the kinase [28–30], demonstrated
51 by number of first generation EGFR inhibitors. Selectivity considerations can also lead to beneficial effects:
52 Imatinib, initially developed to target BCR-Abl fusion proteins, is also approved for treating gastrointestinal
53 stromal tumors (GIST) [31] due to its activity against receptor tyrosine kinase KIT.

54 Use of physical modeling to predict selectivity is relatively unexplored

55 While predicting selectivity is important for drug discovery, but the utility of free energy methods for
56 predicting this property is relatively unexplored. If there is fortuitous cancellation of errors for closely
57 related systems, free energy methods may be much more accurate than expected given the errors made in
58 predicting the potency for each individual target. The selectivity of Imatinib for Abl kinase over Src [32, 33]
59 and within a family of non-receptor tyrosine kinases [34] has been studied extensively using molecular
60 dynamics and free energy calculations. This work focuses on understanding the role reorganization energy
61 plays in the exquisite selectivity of imatinib for Abl over Src despite high similarity between cocrystallized
62 binding mode and kinase conformations, and does not touch on the evaluation of the accuracy of these
63 methods, or their application to drug discovery on congeneric series of ligands. Previous work predicting the
64 selectivity of three bromodomain inhibitors across the bromodomain family achieved promising accuracy
65 for single target potencies of roughly 1 kcal/mol, but does not explicitly evaluate any selectivity metrics [35]
66 or look at correlation in the errors made for each bromodomain.

67 Kinases are an interesting and particularly challenging model system for selectivity predictions
68 Kinases are a useful model system to work with for assessing the utility of free energy calculations to predict
69 selectivity. With the approval of imatinib for the treatment of chronic myelogenous leukemia in 2001,
70 targeted small molecule kinase inhibitors (SMKIs) have become a major class of therapeutics in treating
71 cancer and other diseases. Currently, there are 43 FDA-approved SMKIs [36], and it is estimated that
72 kinase targeted therapies account for as much as 50% of current drug development [37], with many more
73 compounds currently in clinical trials. While there have been a number of successes, the current stable of
74 FDA-approved kinase inhibitors targets only a small number of kinases implicated in disease, and the design
75 of new selective kinase inhibitors remains a significant challenge. Achieving desired selectivity profiles is
76 particularly difficult for kinase targets, making them a system where physical modelling has the potential for
77 a large impact. Achieving selective inhibition of kinases is challenging as there are more than 518 protein
78 kinases [38, 39] with a highly conserved ATP binding site that is targeted by the majority of SMKIs [40]. While
79 kinase inhibitors have been designed to target kinase-specific subpockets and binding modes to achieve
80 selectivity [41–46], previous work has shown that both Type I (binding to the active, DFG-in conformation) and
81 Type II (binding to the inactive, DFG-out conformation) inhibitors display a wide variety of selectivities [47, 48],
82 often exhibiting significant binding to a number of other targets in addition to their primary target. Even
83 FDA-approved inhibitors—often the result of extensive drug development programs—bind to a large number
84 of off-target kinases [49]. Kinases are also targets of interest for developing polypharmacological compounds,
85 or inhibitors that are specifically designed to inhibit multiple kinase targets. Resistance to MEK inhibitors
86 in KRAS-mutant lung and colon cancer has been shown to be driven by HER3 upregulation [50], providing
87 rational for dual MEK/ERBB family inhibitors. Similarly, combined MEK and VEGFR1 inhibition has been
88 proposed as a combinatorial approach to treat KRAS-mutant lung cancer [51]. Developing inhibitors with
89 the desired polypharmacology means navigating more complex selectivity profiles. In well-behaved kinase
90 systems, free energy calculations potency predictions have achieved mean unsigned errors of less than 1.0
91 kcal/mol [7, 12], suggesting that kinases can be computationally tractable as well as clinically interesting.

92 Assessing the ability of alchemical free energy methods to predict selectivity
 93 We anticipate difficulty in predicting selectivity if the errors in the alchemical free energy calculations for two
 94 targets are largely uncorrelated, or even anticorrelated. However, correlation in the forcefield errors of the
 95 free energies for the two targets could lead to a fortuitous cancellation of errors in predicting the selectivity
 96 between targets, making selectivity predictions *more* accurate than potency predictions. Such correlation
 97 could occur because the same chemical elements appear in the ligand and in highly related binding sites.
 98 Here, we investigate the magnitude of this correlation (ρ) and the utility of alchemical free energy calculations
 99 for the prediction of selectivity, hereafter taken to mean the $\Delta\Delta G$ in binding free energies of the same
 100 compound for two targets. We employed state of the art relative free energy calculations [12, 13] to predict
 101 the selectivities of two different congeneric ligand series [52, 53], as well as present a simple numerical
 102 model to quantify the potential speed up in selectivity optimization expected for different combinations
 103 of per target errors and correlation coefficient values. To tease out the effects of a limited number of
 104 experimental measurements, we develop a new Bayesian approach to quantify the uncertainty in the
 105 correlation coefficient in the predicted change in selectivity on ligand modification, incorporating all sources
 106 of uncertainty and correlation in the computation to separate statistical from force field error. We find
 107 that in the closely related systems of CDK2 and CDK9, a high correlation of force field errors suggests that
 108 free energy methods can have a significant impact on speeding up selectivity optimization. In the more
 109 distantly related case (CDK2/ERK2), limited correlation hampers the ability for free energy methods to speed
 110 up selectivity optimization.

111 Methods

112 Numerical model of selectivity optimization speedup

113 To model the impact correlation of forcefield error would have on the expected uncertainty for selectivity pre-
 114 dictions in the regime of infinite sampling and zero statistical error, $\sigma_{selectivity}$ was calculated using Equation 1
 115 for 1000 evenly spaced values of the correlation coefficient (ρ) from 0 to 1, for a number of combinations of
 116 per target forcefield errors ($\sigma_{ff,1}$ and $\sigma_{ff,2}$)

$$\sigma_{selectivity} = \sqrt{\sigma_{ff,1}^2 + \sigma_{ff,2}^2 - 2\rho\sigma_{ff,1}\sigma_{ff,2}} \quad (1)$$

117 The speed up in selectivity optimization that could be expected from using free energy calculations
 118 of a particular per target error ($\sigma_{selectivity}$) was quantified as follows using NumPy (v 1.14.2). An original,
 119 true distribution for the change in selectivity of 200000000 new compounds proposed with respect to a
 120 reference compound was modeled as a normal distribution centered around 0 with a standard deviation of 1
 121 kcal/mol. This assumption was made on the basis that the majority of selectivity is driven by the scaffold, and
 122 R group modifications will do little to drive changes in selectivity. The 1 kcal/mol distribution is supported by
 123 the standard deviations of the selectivity in the experimental datasets referenced in this work, which are all
 124 less than, but close, to 1 kcal/mol.

125 Each of these proposed compounds were "screened" by a free energy calculation technique with a per
 126 target forcefield error (σ_{ff}) of 1 kcal/mol and a specified correlation coefficient ρ . A $\sigma_{selectivity}$ was calculated
 127 according to Equation 1. The noise of the computational method was modeled as a normal distribution
 128 centered around 0 with a standard deviation of $\sigma_{selectivity}$ and added to the "true" change in selectivity, giving
 129 us the predicted change in selectivity ($\Delta S_{compound}$). This process can be described by Equation 2:

$$\Delta S_{compound} = \mathcal{N}_{true}(\mu = 0, \sigma^2 = 1) + \mathcal{N}_{forcefield}(\mu = 0, \sigma_{selectivity}^2(\rho)) \quad (2)$$

130 Any compound predicted to have an improvement in selectivity of 1.4 kcal/mol (1 log unit) would then be
 131 made and have its selectivity experimental measured, using an experimental method with perfect accuracy.
 132 The speedup value for each value of ρ is calculated as the proportion of compounds made with a true
 133 selectivity gain of 1.4 kcal/mol divided by the proportion of compounds with a 1.4 kcal/mol improvement in
 134 the original distribution, where all of the compounds were made.

135 This process was repeated for a 100x (2.8 kcal/mol, 2 log unit) selectivity optimization and 50 linearly
 136 spaced values of the correlation coefficient (ρ) between 0 and 1, for four values of $\sigma_{selectivity}$ and 40000000
 137 compounds in the original distribution.

138 Numerical model of impact of statistical error on selectivity optimization
 139 To model the impact of statistical error on selectivity optimization at different levels of correlation in the
 140 forcefield error, a similar scheme as above was used. An original, true distribution of 40000000 compounds
 141 was proposed with respect to a reference compound, drawing from a normal distribution centered on 0
 142 with a standard deviation of 1 (Numpy v 1.14.2). Each of these proposed compounds were "screened" by a
 143 free energy calculation technique with a per target forcefield error (σ_{ff}) of 0.9 kcal/mol [57] and a specified
 144 correlation coefficient ρ , which was evenly spaced between 0 and 1 in 50 steps. A $\sigma_{selectivity}$ was calculated
 145 according to Equation 1. Additionally, a per target statistical error (σ_{stat}) was as in Equation 3

$$\sigma_{stat} = \sqrt{\frac{2\sigma^2}{N}} \quad (3)$$

146 Where N is the effort put into running sampling the calculation and σ is such that when N is 1, $\sigma_{stat} = 0.2$
 147 kcal/mol. The statistical error was propagated assuming it was uncorrelated, such as in Equation 1 where $\rho =$
 148 0, giving us $\sigma_{statistics}$. The forcefield and statistical errors were modeled as Gaussian noise added to the true
 149 distribution, as in Equation 4.

$$\Delta S_{compound} = \mathcal{N}_{true}(\mu = 0, \sigma^2 = 1) + \mathcal{N}_{forcefield}(\mu = 0, \sigma_{selectivity}^2(\rho)) + \mathcal{N}_{stat}(\mu = 0, \sigma_{statistics}^2(\rho)) \quad (4)$$

150 Any compound predicted to have an improvement in selectivity of 2.8 kcal/mol (2 log units) would then be
 151 made and have its selectivity experimental measured, using an experimental method with perfect accuracy.
 152 The speedup value for each value of ρ is calculated as the proportion of compounds made with a true
 153 selectivity gain of 2.8 kcal/mol divided by the proportion of compounds with a 2.8 kcal/mol improvement in
 154 the original distribution, where all of the compounds were made.

155 Structure Preparation

156 Structures from the Shao [52] and Hole [54], and Blake [53] papers were downloaded from the PDB [55], selecting
 157 structures with the same co-ligand crystallized. For the Shao dataset, 4BCK (CDK2) and 4BCI (CDK9) were
 158 selected, which have ligand 12c cocrystallized. For the Blake dataset, 5K4J (CDK2) and 5K4I (ERK2) were
 159 selected, cocrystallized with ligand 21. The structures were prepared using Schrodinger's Protein Preparation
 160 Wizard [56] (release 2017-3). This pipeline modeled in internal loops and missing atoms, added hydrogens at
 161 the reported experimental pH (7.0 for the Shao dataset, 7.3 for the Blake dataset) for both the protein and
 162 the ligand. All crystal waters were retained. The ligand was assigned protonation and tautomer states using
 163 Epik at the experimental pH \pm 2, and hydrogen bonding was optimized using PROPKA at the experimental
 164 pH \pm 2. Finally, the entire structure was minimized using OPLS3 with an RMSD cutoff of 0.3Å.

165 Ligand Pose Generation

166 Ligands were extracted from the publication entries in the BindingDB as 2D SMILES strings. 3D conformations
 167 were generated using LigPrep with OPLS3 [57]. Ionization state was assigned using Epik at experimental
 168 pH \pm 2. Stereoisomers were computed by retaining any specified chiralities and varying the rest. The tautomer
 169 and ionization state with the lowest epik state penalty was selected for use in the calculation. Any ligands
 170 predicted to have a positive or negative charge in its lowest Epik state penalty was excluded, with the
 171 exception of Compound 9 from the Blake dataset. This ligand was predicted to have a +1 charge for its
 172 lowest state penalty state. The neutral form the ligand was include for the sake of cycle closure in the FEP+
 173 map, but was ignored for the sake any analysis afterwards. Ligand poses were generated by first aligning to
 174 the co-crystal ligand using the Largest Common Bemis-Murcko scaffold with fuzzy matching (Schrodinger
 175 2017-4). Ligands that were poorly aligned or failed to align were then aligned using Maximum Common
 176 Substructure (MCSS). Finally, large R-groups were allowed to sample different conformations using MM-GBSA
 177 with a common core restrained. VSGB solvation model was used with the OPLS3 forcefield. No flexible
 178 residues were defined for the ligand.

179 Free Energy Calculations

180 The FEP+ panel (Maestro release 2017-4) was used to generate perturbation maps. FEP+ calculations were
 181 run using the FEP+ panel from Maestro release 2018-3, using the parameters from the version of OPLS3e
 182 that shipped with the 2018-3 release. Any missing ligand torsions were fit using the automated FFbuilder
 183 protocol [58]. Custom charges were assigned using the OPLS3e forcefield using input geometries, according
 184 to the automated FEP+ workflow released in 2018-3. Neutral perturbations were run for 15ns per replica,
 185 using an NPT ensemble and water buffer size of 5Å. The SPC water model was used. A GCMC solvation
 186 protocol was used to sample buried water molecules in the binding pocket prior to the calculation, which
 187 discards any retained crystal waters.

188 Statistical Analysis of FEP+ calculations

189 Each FEP+ calculation has a reported mean unsigned error (MUE) and root mean squared error (RMSE) with
 190 a bootstrapped 95% confidence interval. The MUE was calculated according Equation 5, while the RMSE was
 191 calculated according to Equation 6.

$$MUE = \frac{\sum_0^n |\Delta G_{calc} - \Delta G_{exp}|}{n} \quad (5)$$

$$RMSE = \frac{\sum_0^n \sqrt{\Delta G_{calc}^2 - \Delta G_{exp}^2}}{n} \quad (6)$$

192 Each RMSE and MUE is reported with a 95% confidence interval calculated from 10000 replicates of a
 193 choose-one-replace bootstrap protocol on the ΔG values reported to account for the finite sample size of the
 194 ligands. The code used to bootstrap these values is available on github: <https://github.com/choderlab/selectivity>

195 Quantification of the correlation coefficient ρ

196 To quantify ρ , we built a Bayesian graphical model using pymc3 (v. 3.5) [60] and theano (v 1.0.3) [61], which
 197 has been made available on Github. For each phase (complex and solvent), the absolute free energy (G)
 198 of ligand i was treated as a normal distribution (Equation 7). For each set of calculations, one ligand was
 199 chosen as the reference, and pinned to 0, with a standard deviation of 1 kcal/mol in order to improve the
 200 efficiency of sampling from the model.

$$G_{i,target}^{phase} = \mathcal{N}(\mu = 0, sd = 25.0 \text{ kcal/mol}) \quad (7)$$

201 For each edge of the FEP map (ligand $i \rightarrow$ ligand j), there is a contribution from dummy atoms, that was
 202 modeled as in Equation 8.

$$c_{i,j} = \mathcal{N}(\mu = 0, sd = 25.0 \text{ kcal/mol}) \quad (8)$$

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

$$\Delta G_{phase, ij, target}^{BAR} = \mathcal{N}(G_{j,target}^{phase} - G_{i,target}^{phase}, \delta^2 \Delta G_{phase, ij, target}^{BAR}, observed = \Delta G_{phase, ij, target}^{calc}) \quad (9)$$

204 Where $\delta^2 \Delta G_{phase, ij, target}^{BAR}$ is the reported BAR uncertainty from the calculation, and $\Delta G_{phase, ij, target}^{calc}$ is the
 205 BAR estimate of the free energy for the perturbation between ligands i and j in a given phase.

206 From this, we can calculate the $\Delta\Delta G^{FEP}$ for each edge as in Equation 10:

$$\Delta\Delta G_{target, ij}^{FEP} = \Delta G_{complex, ij, target}^{BAR} - \Delta G_{solvent, ij, target}^{BAR} \quad (10)$$

207 To model the way an offset is calculated for the ΔG reported by the FEP+ panel in Maestro:

$$\text{offset} = \frac{\sum^n G_{i,target}^{complex} - G_{i,target}^{solvent}}{n} - \frac{\sum^n \Delta G_i^{exp}}{n} \quad (11)$$

208 The offset was added to each ΔG_i^{BAR} to calculate ΔG_i^{sch} .

209 The experimental binding affinity was treated as a true value ($\Delta G_{i,target}^{true}$) corrupted by experimental
 210 uncertainty, which is assumed to be 0.3 kcal/mol [6], with the values reported in the papers ($\Delta G_{i,target}^{obs}$) treated
 211 as observations from this distribution (Equation 12)

$$\Delta G_{i,target}^{exp} = \mathcal{N}(\Delta G_{i,target}^{true}, 0.3 \text{ kcal/mol}, observed = \Delta G_{i,target}^{obs}) \quad (12)$$

212 $\Delta G_{i,target}^{true}$ was assigned a weak normal prior, as in equation 13.

$$\Delta G_{i,target}^{true} = \mathcal{N}(0, 50 \text{ kcal/mol}) \quad (13)$$

213 The error for a given ligand was calculated as in Equation 14.

$$\epsilon_i = \Delta G_i^{sch} - \Delta G_i^{true} \quad (14)$$

214 From these ϵ values, we calculated the correlation coefficient, ρ as in Equation 15.

$$\rho = \frac{cov(\epsilon_{target1}, \epsilon_{target2})}{\sigma_{target1}\sigma_{target2}} \quad (15)$$

215 Where σ is the standard deviation of ϵ . To quantify ρ for the CDK2/ERK2 calculations, the default NUTS
 216 sampler with jitter+adapt_diag initialization, 1000 tuning steps, and a target accept probability of 0.8 was
 217 used to draw 10000 samples from the model. The CDK2/CDK9 model was sampled 20000 times using default
 218 NUTS sampler with jitter+adapt_diag initialization and 3000 tuning steps.

219 Calculating the marginal distribution of speedup

220 To quantify the expected speedup from the calculations we ran, we utilized 10000 replicates of the scheme
 221 detailed above to calculate speedup given parameters ρ , $\sigma_{ff,1}$ and $\sigma_{ff,2}$, in the regime of infinite effort
 222 and zero statistical error. Using Numpy (v 1.14.2), ρ was drawn from a normal distribution with the mean
 223 and standard deviation from the posterior distribution of ρ from the Bayesian Graphical model. The per
 224 target forcefield errors, $\sigma_{ff,1}$ and $\sigma_{ff,2}$, were estimated from the mean of the absolute value of $\epsilon_{target1}$ and
 225 $\epsilon_{target2}$, which are the magnitude of errors from the Bayesian graphical model. $\sigma_{selectivity}$ was calculated using
 226 Equation 1. 100000 molecules were proposed from true normal distribution, as above. The error of the
 227 computational method was modeled as in Equation 2.

228 Results

229 Free energy methods can be used to predict the selectivity of a compound
 230 While ligand potency for a single target is often quantified as a free energy of binding ($\Delta G_{binding}$), there are
 231 a number of different metrics for quantifying the selectivity of a compound [62, 63]. Here, we propose
 232 a more granular view of selectivity: the change in free energy of binding for a given ligand between two
 233 different targets ($\Delta\Delta G_{selectivity}$), which can be calculated as in Equation 16. $\Delta\Delta G_{selectivity}$ is a useful measure of
 234 compound selectivity once a single, or small panel, of off-targets have been identified.

$$\Delta\Delta G_{selectivity} = \Delta G_{binding, target 2} - \Delta G_{binding, target 1} \quad (16)$$

235 To predict the $\Delta\Delta G_{selectivity}$ of a compound, we developed a protocol that uses a relative free energy
 236 calculation (FEP+) [12] to run a map of perturbations between ligands in a congeneric series, as described
 237 in depth in the methods section. The calculation is repeated for each target of interest, with identical
 238 perturbations (edges) between each ligand (nodes). Each edge represents a relative free energy calculation
 239 that quantifies the $\Delta\Delta G$ between the ligands, or nodes. By using provided experimental data, we can convert
 240 the $\Delta\Delta G$ from each edge to a single potency value for each value against that target (ΔG_{target}). From this
 241 sets of calculations, we can calculate a $\Delta\Delta G_{selectivity}$ for each ligand given two targets of interest. Previous
 242 work shows that FEP+ can achieve an accuracy (σ_{target}) of roughly 1 kcal/mol when predicting potency, which
 243 is a combination of systematic forcefield and random statistical error [12]. However, it is possible that the
 244 forcefield component of that error (σ_{ff}) may fortuitously cancel when computing $\Delta\Delta G_{selectivity}$, leading to a
 245 forcefield component of the selectivity uncertainty that is lower than would be expected.

246 Correlation of errors can make selectivity predictions more accurate and speed up ligand optimi-
 247 zation
 248 To demonstrate the potential impact correlation has on the forcefield error of selectivity predictions ($\sigma_{selectivity}$)
 249 using alchemical free energy techniques, we created a simple numerical model following equation 1, which
 250 takes into account each of the per target forcefield errors expected from the methodology as well as the
 251 correlation in those errors. As seen in Figure 1A, if the per target forcefield errors ($\sigma_{ff,1}$ and $\sigma_{ff,2}$) are
 252 the same, $\sigma_{selectivity}$ approaches 0 as the correlation coefficient (ρ) approaches 1. If the error for the free
 253 energy method is not the same, $\sigma_{selectivity}$ gets smaller but approaches a non-zero value as ρ approaches 1. To
 254 quantify the expected speedup in selectivity optimization, we modeled the change in selectivity with respect
 255 to a reference compound for a number of compounds a medicinal chemist might suggest as a normal
 256 distribution centered around 0 with a standard deviation of 1 kcal/mol (Figure 1B, black curve), reflecting
 257 that most proposed modifications would not drive large changes in selectivity. Then, suppose that each
 258 compound is screened computationally with a method free energy methodology with a per target forcefield
 259 error (σ_{ff}) of 1 kcal/mol in the regime of infinite computation effort where statistical error is 0 kcal/mol. All
 260 compounds predicted to have a 1.4 kcal/mol improvement in selectivity are synthesized and experimentally
 261 tested (Figure 1B, colored curves), using an experimental technique with perfect accuracy. The fold-change
 262 in the proportion of compounds that are made that have a true 1.4 kcal/mol improvement in selectivity
 263 compared to the original distribution can be calculated as a surrogate for the expected speedup. For a 1.4
 264 kcal/mol selectivity improvement threshold (1 log unit), a correlation of 0.5 gives an expected speed up of
 265 4.1x, which can be interpreted as 4.1x fewer compounds needing to be made before achieving a 1 log unit
 266 improvement in selectivity. This process can be extended for the even more difficult proposition of achieving
 267 a 2 log unit improvement in selectivity (Figure 1C), where 200-300x speedups can be expected, depending on
 268 σ_{ff} for the free energy methodology.

269 The CDK2 and CDK9 experimental dataset demonstrates the difficulty in achieving selectivity for
 270 closely related kinases

271 I will insert some sentences about the relatedness of the kinases based on Andrea's analysis
 272
 273 To begin quantifying the correlation of errors in free energy predictions for selectivity, we set out to gather
 274 datasets that met a number of criteria. We looked for datasets that contained binding affinity data for a
 275 number of kinase targets and ligands, as well as having crystal structures for each target with the same
 276 co-crystallized ligand. For the CDK2/CDK9 dataset [52], ligand 12c was cocrystallized with CDK2/cyclin A
 277 (Figure 2A, left) and CDK9/cyclin T (Figure 2B, left), work that was published in a companion paper [54]. In
 278 both CDK2 and CDK9, ligand 12c forms relatively few hydrogen bond interactions with the kinase. Each
 279 kinase forms a set of hydrogen bonds between the ligand scaffold and a hinge residue (C106 in CDK9 and
 280 L83 in CDK2) that is conserved across all of the ligands in this series. CDK9, which has slightly lower affinity
 281 for ligand 12c (Figure 2C, right), forms a lone interaction between the sulfonamide of ligand 12c and residue
 282 E107. On the other hand, CDK2 forms interactions between the sulfonamide of ligand 12c and residues
 283 K89 and H84. The congeneric series of ligands contains a number of challenging perturbations, particularly
 284 at substituent point R3 (Figure 2C, left). Ligand 12i also presented a challenging perturbation, moving the
 285 1-(piperazine-1-yl)ethanone from the *meta* to *para* location.

286 This congeneric series of ligands also highlights two of the challenges of working from publicly available
 287 data. First, the dynamic range of selectivity is incredibly narrow, with a mean $\Delta\Delta G_{selectivity}$ (CDK9 - CDK2)
 288 of only -0.65 kcal/mol, and a standard deviation of 0.88 kcal/mol. Additionally, experimental uncertainties
 289 are not reported for the experimental measurements. Thus, for this and subsequent sets of ligands,
 the experimental uncertainty is assumed to be 0.3 kcal/mol based on previous work done to summarize
 uncertainty in experimental data [6, 64].

290 The CDK2 and ERK2 dataset achieves higher levels of selectivity for more distantly related kinases
 291 The CDK2/ERK2 dataset from Blake *et al.*, 2016 also met the criteria described above. Crystal structures
 292 for both CDK2 (Figure 3A, top) and ERK2 (Figure 3B, top) were available with ligand 22 co-crystallized. Of
 293 note, CDK2 was not crystallized with cyclin A, despite cyclin A being included in the affinity assay reported
 294 in the paper [53]. CDK2 adopts a DFG-in conformation with the α C helix rotated out, away from the ATP

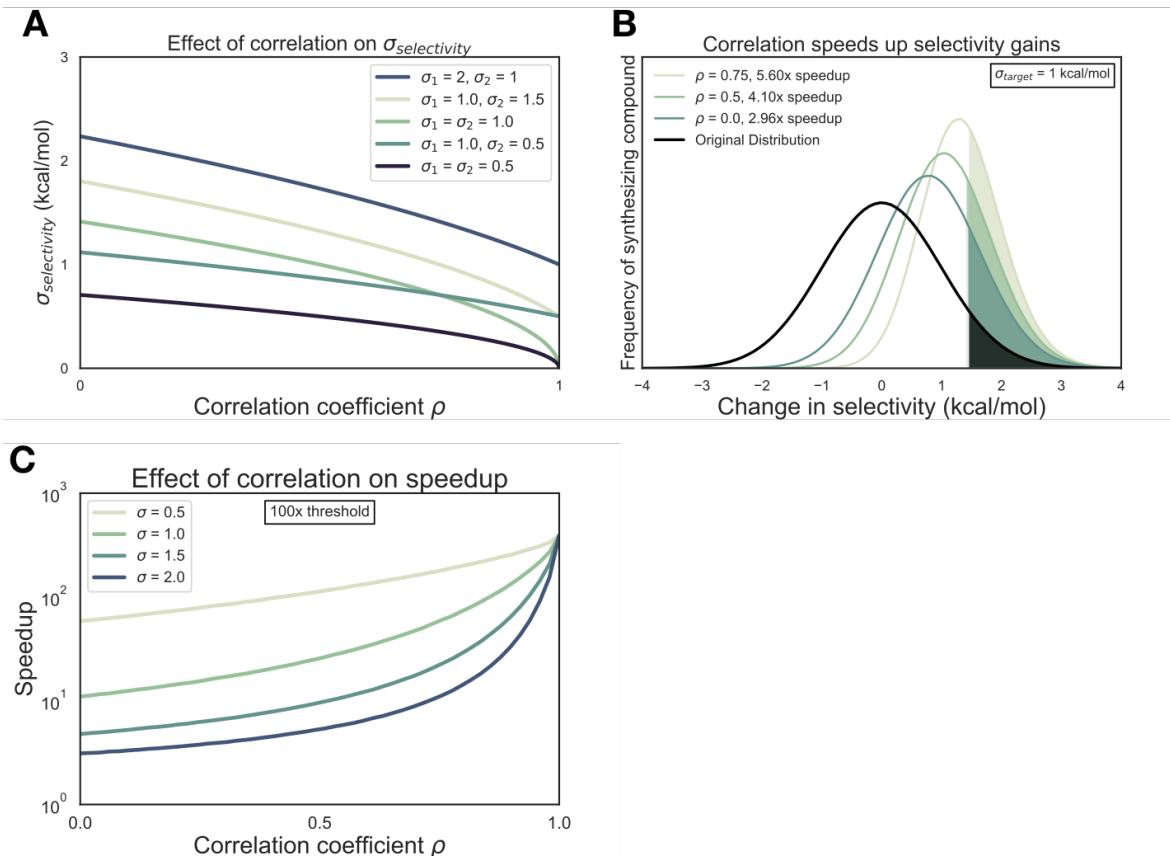
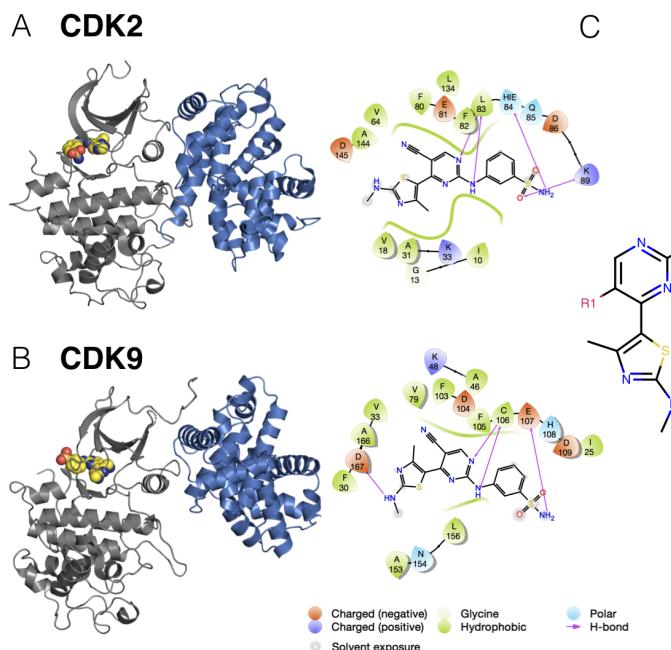


Figure 1. Free energy calculations speed up selectivity optimization

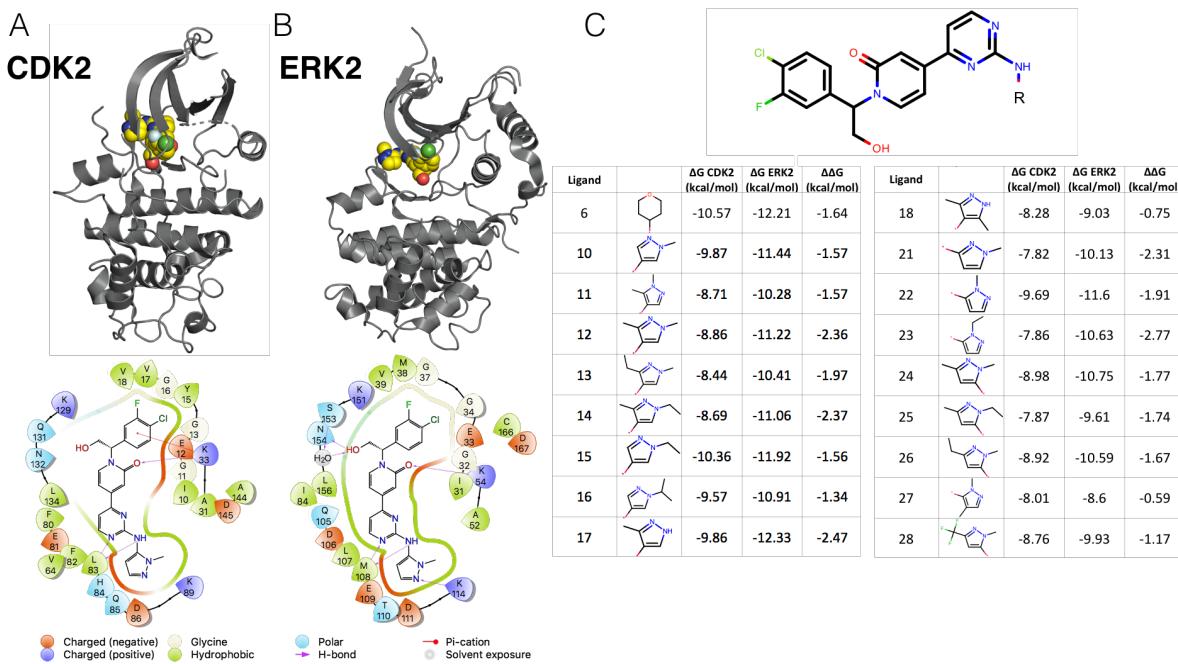
(A) The effect of correlation on expected errors for predicting selectivity ($\sigma_{selectivity}$) in kcal/mol. Each curve represents a different combination of per target forcefield errors ($\sigma_{ff,1}$ and $\sigma_{ff,2}$). (B) The change in selectivity for molecules proposed by medicinal chemists optimizing a lead candidate can be modeled by a normal distribution centered on 0 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 unit (1.4 kcal/mol) improvement in selectivity with a free energy methodology with a 1 kcal/mol per target forcefield error and a particular correlation, in the regime of infinite error where statistical error is zero. The shaded region of each curve corresponds to the compounds with a real 1 log unit improvement in selectivity. The speedup is calculated as the ratio of the percentage of compounds made with a real 1 log unit improvement to the percentage of compounds that would be expected in the original distribution. (C) The speedup (y-axis, log scale) expected for 100x (2 log units, 2.8 kcal/mol) selectivity optimization as a function of correlation coefficient ρ . Each curve corresponds to a different σ_{ff} value.



Ligand	R1	R2	R3	ΔG CDK2 (kcal/mol)	ΔG CDK9 (kcal/mol)	$\Delta\Delta G$ (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 selectivity dataset from Shao et al., 2013

(A) (left) Crystal Structure (4BCK)[54] 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)[54] (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[52].

**Figure 3. CDK2 and ERK2 selectivity dataset from Blake et al., 2016**

(A) (top) Crystal structure of CDK2 (5K4J) shown in gray cartoon and ligand 22 shown in yellow spheres. (bot) 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. (bot) 2D interaction map of ligand 22 in the binding pocket of ERK2. **(C)** (top) Common scaffold for all of the ligands in the Blake dataset, with R denoting attachment side for substitutions. (bot) Table showing R group substitutions and experimentally measured binding affinities and selectivities. Ligand numbers correspond to those used in publication.

295 binding site and breaking the conserved salt bridge between K33 and E51 (Supp. Figure 1A), indicative of an
296 inactive kinase [43, 65]. By comparison, the CDK2 structure from the CDK2/CDK9 dataset adopts a DFG-in
297 conformation with the α C helix rotated in, forming the ionic bond between K33 and E51 indicative of an
298 active kinase, due to allosteric activation by cyclin A. While missing cyclins have caused problems for free
299 energy calculations in prior work, it is possible that the fully active conformation contributes equally to
300 binding affinity for all of the ligands in the series, and the high accuracy of the potency predictions (Figure 4,
301 top left) is the result of fortuitous cancellation of errors. The binding mode for this series is similar between
302 both kinases. There is a set of conserved hydrogens bonds between the scaffold of the ligand and the
303 backbone of one of the hinge residues (L83 for CDK2 and M108 for ERK2). The conserved lysine (K33 for
304 CDK2 and K54 for ERK2), normally involved in the formation of a ionic bond with the α C helix, forms a
305 hydrogen bond with the scaffold (Figure 4A and 4B, bottom) in both CDK2 and ERK2. However, in the ERK2
306 structure, the hydroxyl engages a crystallographic water as well as N154 in a hydrogen bond network that is
307 not present in the CDK2 structure. The congenic ligand series features a single substituent point, with the R
308 groups exposed to the solvent. This helps explain the extremely narrow distribution of selectivities, with a
309 mean selectivity of -1.74 kcal/mol (ERK2 - CDK2) and standard deviation of 0.56 kcal/mol. This suggests that
310 the selectivity is largely driven by the scaffold and unaffected by the R group substitutions.

311 FEP+ calculations show accurate potency predictions for ERK2/CDK2 and larger errors for CDK2/CDK9
312 The FEP+ predictions of single target potencies (ΔG) showed good accuracy for the CDK2 and ERK2 dataset
313 (Figure 4, top). Replicate 1 of the calculations in shown in Figure 4, with an RMSE of $0.41^{0.71}_{0.13}$ and $0.79^{1.48}_{0.09}$
314 kcal/mol, respectively. All of the CDK2 potencies were predicted within 1 log unit of the experimental value,
315 while ERK2 had two outliers. The selectivity ($\Delta G_{selectivity}$) predictions show an RMSE of $0.77^{1.38}_{0.22}$ kcal/mol, with
316 all but one of the predictions falling within 1 log unit of the experimental values (Figure 4, top right panel).

317 This was consistent across all three replicates of the calculations (Supp. Figure 6). This consistency across
 318 replicates holds true at the individual ligand level as well (Supp. Figure 8). Despite the low RMSE for the
 319 selectivity predictions, the narrow dynamic range and high uncertainty from experiment and calculation
 320 makes it difficult to determine which compounds are more selective than others.

321 Replicate 1 of the CDK2/CDK9 calculations are shown in the bottom panel of Figure 4. The CDK2 and
 322 CDK9 datasets show higher errors in the potency predictions, with an RMSE of $1.04^{1.86}_{0.27}$ and $1.79^{2.65}_{0.84}$ kcal/mol
 323 respectively. There are a number of outliers that fall outside of 1 log unit from the experimental value for
 324 CDK2 and CDK9. While the higher per target errors make predicting potency more difficult, the selectivity
 325 predictions show a much lower RMSE of $0.70^{1.06}_{0.30}$ kcal/mol. This suggests that some correlation in the error is
 326 leading to fortuitous cancellation of the forcefield error, leading to more accurate than expected predictions
 327 of $\Delta\Delta G_{selectivity}$. These results were consistent across all three replicates of the calculation (Supp. Figure 4) as
 328 well as each individual ligand (Supp. Figure 8).

329 Correlation of forcefield errors accelerates selectivity optimization

330 To quantify the correlation coefficient (ρ) of the forcefield errors in our calculations, we built a Bayesian
 331 graphical model to separate the forcefield error from the statistical error, as described in depth in the
 332 methods section. Briefly, we modeled the absolute free energy (G) of each ligand in each phase (complex and
 333 solvent) as in equation 7. The model was chained to the FEP+ calculations by providing the $\Delta G_{phase,ij,target}^{calc}$ as
 334 observed data, as in equation 9. As in equation 10, the experimental data was modeled as a normal distribution
 335 centered around the true free energy of binding ($\Delta G_{i,target}^{true}$) corrupted by experimental error, which is assumed
 336 to be 0.3 kcal/mol from previous work done to quantify the uncertainty in publicly available data [6]. The
 337 reported IC50 values from each dataset were treated as data observations (Equation 12) and the $\Delta G_{i,target}^{true}$
 338 was assigned a weak normal prior (Equation 13). The correlation coefficient was calculated for each sample
 339 according to equation 14. The correlation coefficient ρ for replicate 1 of the CDK2/ERK2 calculations was
 340 quantified to be $0.49^{0.68}_{0.27}$, indicating that the errors are correlated between ERK2 and CDK2 (Figure 6A, right),
 341 which was consistent with the distributions for ρ in replicates 2 and 3 (Supp. Figure 7). The joint marginal
 342 distribution of the error (ϵ) for each target is more diagonal than symmetric, which is expected for cases in
 343 which ρ is 0.5 (Supp. Figure 2). In addition to correlation in the forcefield errors, the high per target accuracy
 344 of these calculations allow for a predicted 2-3x speed up for 1 log unit selectivity optimization, and a 20-50x
 345 speed up for 2 log unit selectivity optimization (Figure 6A, right), in the regime of infinite sampling effort
 346 where statistical error is 0.

347 The CDK2/CDK9 calculations show strong evidence of correlation, with a correlation coefficient of $0.72^{0.83}_{0.58}$
 348 (Figure 6B, right) for replicate 1. The rest of the replicates showed strong agreement (Supp. Figure 5). The joint
 349 marginal distribution of errors is strongly diagonal, which is expected based on the value for ρ (Figure 6B,
 350 left). The high correlation in errors leads to a speed up of 2-3 for 1 log unit selectivity optimization and
 351 30-40x for 2 log unit selectivity optimization (Figure 6B, right), despite the much higher per target RMSE than
 352 the CDK2/ERK2 case.

353 Quantifying ρ for these calculations enables estimation of the forcefield error in the selectivity predictions,
 354 $\sigma_{selectivity}$. This is useful for estimating expected error for prospective studies, where the experimental values
 355 for $\Delta\Delta G_{selectivity}$ are not yet known. Based on the distribution quantified for ρ , the expected $\sigma_{selectivity}$ for
 356 the CDK2/CDK9 calculations is $1.18^{1.38}_{0.95}$ kcal/mol (Supp. Figure 3), which is in good agreement with the
 357 bootstrapped RMSE (Figure 4, bottom). For the CDK2/ERK2 calculations, $\sigma_{selectivity}$ is $0.96^{1.14}_{0.75}$ (Supp. Figure 3),
 358 which is also in good agreement with the bootstrapped RMSE (Figure 4, top).

359 Expending more effort to reduce statistical error can improve selectivity optimization speedups
 360 To this point, we have considered only forcefield error in quantifying the speedup free energy calculations
 361 can enable for selectivity optimization, by assuming we are in the range of infinite sampling, where the
 362 statistical error for each target is reduced to 0 kcal/mol. To begin understanding how statistical error impacts
 363 this speedup, we modified the model of speedup by additional considering the per target statistical error
 364 (σ_{stat}), which we define in Equation 3 such that at the baseline effort, N , σ_{stat} is 0.2 kcal/mol. In this definition,
 365 it takes 4 times the sampling, or effort, to reduce statistical error by a factor of 2. We assume that statistical

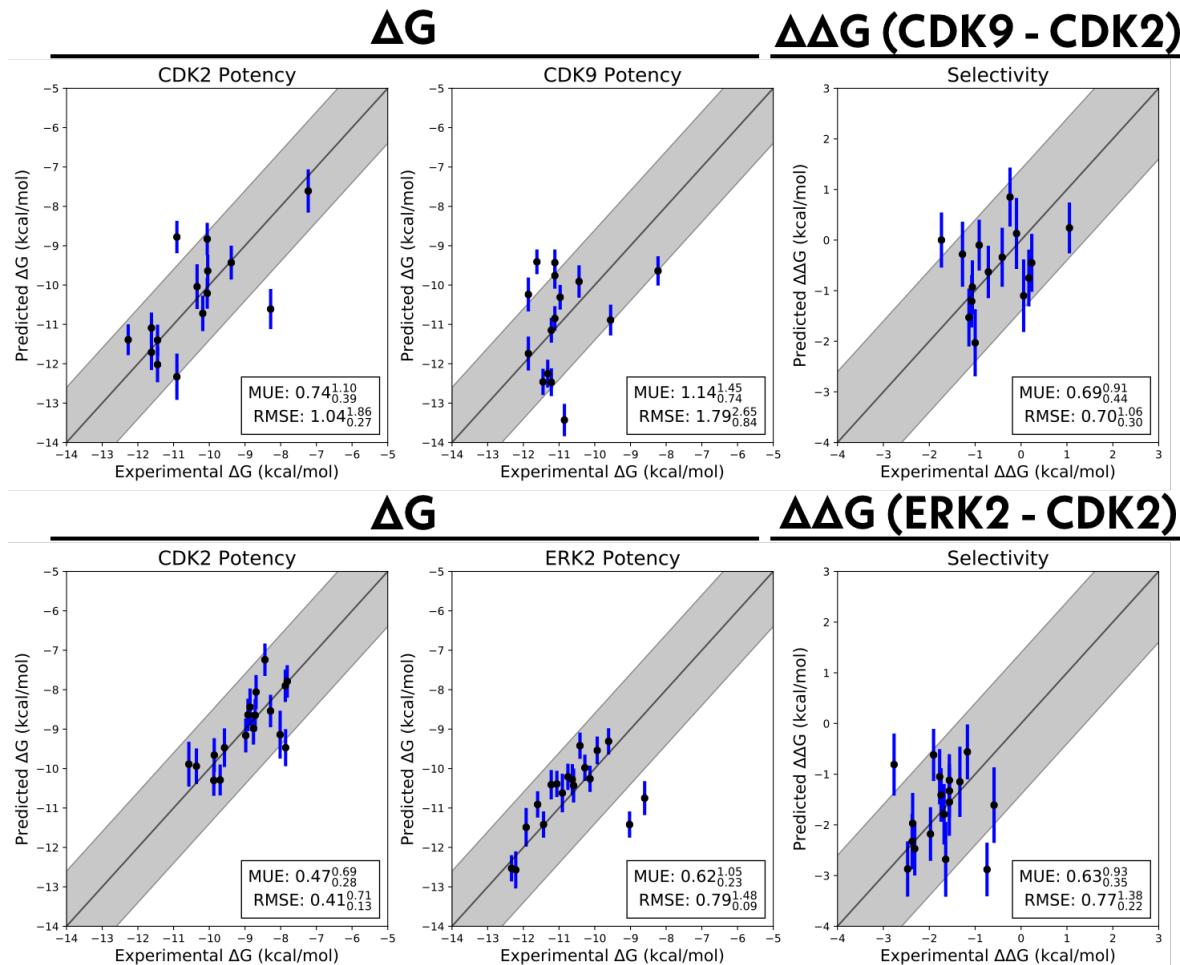


Figure 4. Relative free energy calculations can accurately predict potency, but show larger errors for selectivity predictions.

Single target potencies and selectivities for CDK2/ERK2 from the Blake datasets (*top*), and CDK2/CDK9 (*bottom*) from the Shao datasets. The experimental values are shown on the X-axis and calculated values on the Y-axis. Each data point corresponds to a ligand for a given target. All values are shown in units of kcal/mol. The horizontal error bars show the assumed experimental uncertainty of 0.3 kcal/mol[6]. We show the 95% CI based on the estimated statistical (σ_{stat}) as vertical blue error bars. 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.

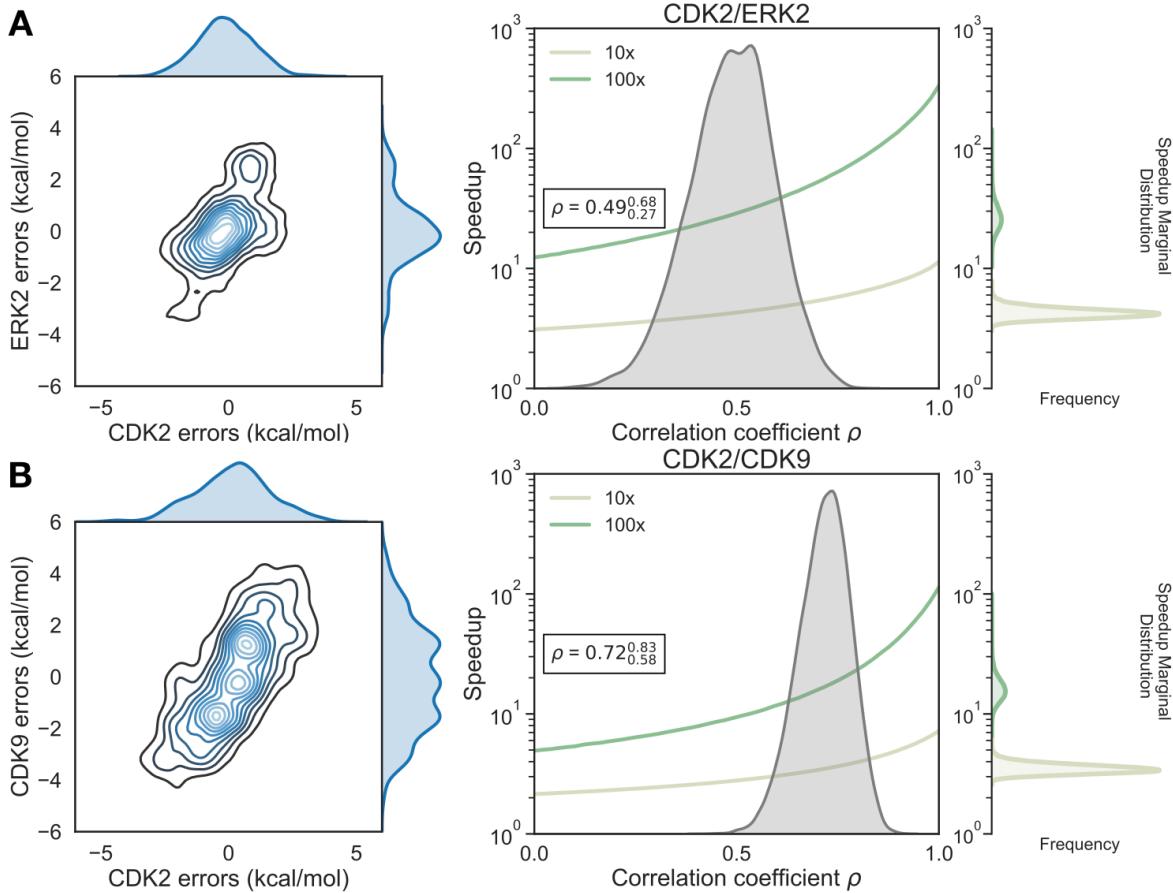


Figure 5. Correlation in selectivity prediction errors can be used to accelerate selectivity optimization

(A) (left) The joint posterior distribution of the prediction errors for CDK2 (X-axis) and ERK2 (Y-axis) from the Bayesian graphical model. (right) Speedup in selectivity optimization (Y-axis) as a function of correlation coefficient (X-axis). The posterior marginal distribution of the correlation coefficient (ρ) is shown in gray, while the expected speed up is shown for 100x (green curve) and 10x (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 100x (green) and 10x (yellow) selectivity optimization speedups. (B) (left) The same as above, with CDK2 (X-axis) and CDK9 (Y-axis). (right) As above, for the CDK2/CDK9 calculations.

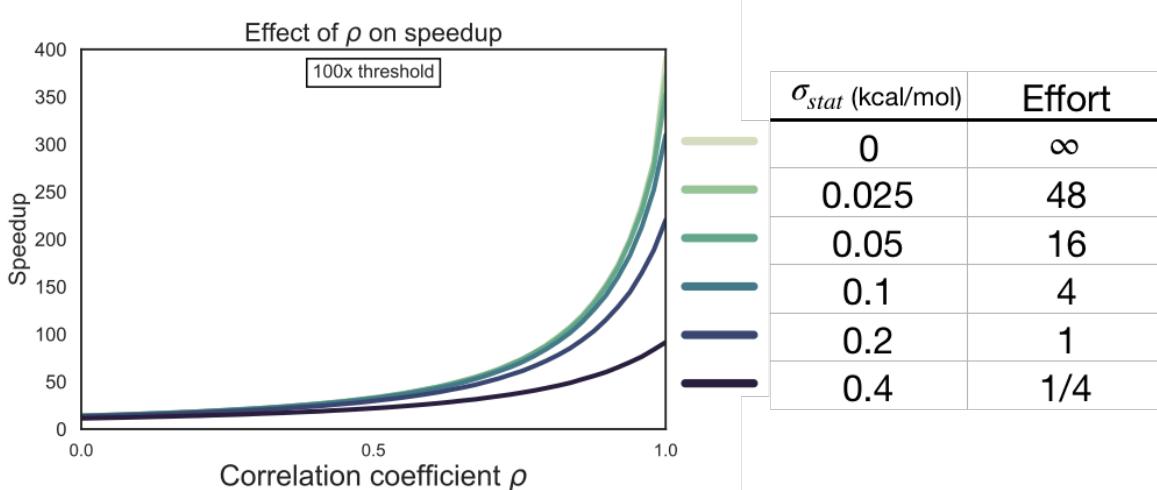


Figure 6. Correlation in selectivity prediction errors can be used to accelerate selectivity optimization

(A) (left) The joint posterior distribution of the prediction errors for CDK2 (X-axis) and ERK2 (Y-axis) from the Bayesian graphical model. (right) Speedup in selectivity optimization (Y-axis) as a function of correlation coefficient (X-axis). The posterior marginal distribution of the correlation coefficient (ρ) is shown in gray, while the expected speed up is shown for 100x (green curve) and 10x (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 100x (green) and 10x (yellow) selectivity optimization speedups. (B) (left) The same as above, with CDK2 (X-axis) and CDK9 (Y-axis). (right) As above, for the CDK2/CDK9 calculations.

366 error is uncorrelated when propagating to two targets, and that σ_{ff} is 0.9 kcal/mol for both targets [57, 64].
 367 As shown in Figure 5, expending effort to reduce σ_{stat} when ρ is less than 0.5 does not change the expected
 368 speedup for the 100x threshold in meaningful way, suggesting that it is not worth running calculations
 369 longer than the default protocol in this case. However, when ρ is greater than 0.5, the curves do start to
 370 separate, particularly the 1/4x, 1x and 4x effort curves. This suggests that when the correlation is high,
 371 running longer calculations can net improvements in selectivity optimization speed. Interestingly, the 16x,
 372 48x and ∞ effort curves do not differ greatly from the 4x effort curve, indicating that there are diminishing
 373 returns to expending more time running longer calculations.

374 In order to understand what the current statistical error is for our calculations, we performed three
 375 replicates of our calculations, and calculated the standard deviation of the cycle closure $\Delta\Delta G$ for each edge
 376 of the map, and compared that value to the cycle closure errors reported for each edge (Supp. Figure 9). In
 377 general, the standard deviation suggests that the statistical error for our calculations is between 0.1 and
 378 0.3 kcal/mol. While this does not agree well with the cycle closure error (Supp. Figure 9), the high variation
 379 of the cycle closure errors between replicates of each edge suggest that the standard deviation is a more
 380 reliable estimate of what the statistical error for these calculations is.

381 Discussion and Conclusions

382 We have demonstrated, using a simple numerical model, the impact that free energy calculations with even
 383 weakly correlated errors can have on speeding up the optimization of selectivity in small molecule kinase
 384 inhibitors. While the expected speed up is dependent on the per target error of the method (σ_{target}), the
 385 speedup is also highly dependent on the correlation of errors made for both targets. Unsurprisingly, free
 386 energy methods have greater impact as the threshold for selectivity optimization goes from 10x to 100x.
 387 While 100x selectivity optimization is difficult to achieve, the expected benefit from free energy calculations
 388 is also quite high, with 1 and 2 order of magnitude speedups possible. To quantify the correlation of errors
 389 in two example systems, we gathered experimental data for two congeneric ligand series with experimental
 390 data for CDK2 and ERK2, as well as CDK2 and CDK9. These datasets, which had crystal structures for both

391 targets with the same ligand co-crystallized, exemplify the difficulty in predicting selectivity. The dynamic
392 range of selectivity for both systems is incredibly narrow, with most of the perturbations not having a major
393 impact on the overall selectivity achieved. Further, the data was reported with unreliable experimental
394 uncertainties, which makes quantifying the errors made by the free energy calculations difficult. This issue is
395 common when considering selectivity, as many kinase-oriented high throughput screens are carried out at a
396 single concentration and not highly quantitative.

397 **Acknowledgments**

398 JDC and SKA acknowledge support from NIH grant R01GM121505. JDC acknowledges partial support from
399 NIH grant P30CA008748. Patrick Grinaway for useful discussions about Bayesian statistics

400 **To be filled out soon**

401 **Disclosures**

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

410 **Author Contributions**

411 Conceptualization: SKA, LW, RA, JDC

412 Methodology: SKA, LW, JDC

413 Investigation: SKA, SP

414 Writing – Original Draft: SKA

415 Writing – Review & Editing: SKA, JDC

416 Funding Acquisition: RA, JDC

417 Resources: LW, JDC

418 Supervision: LW, JDC

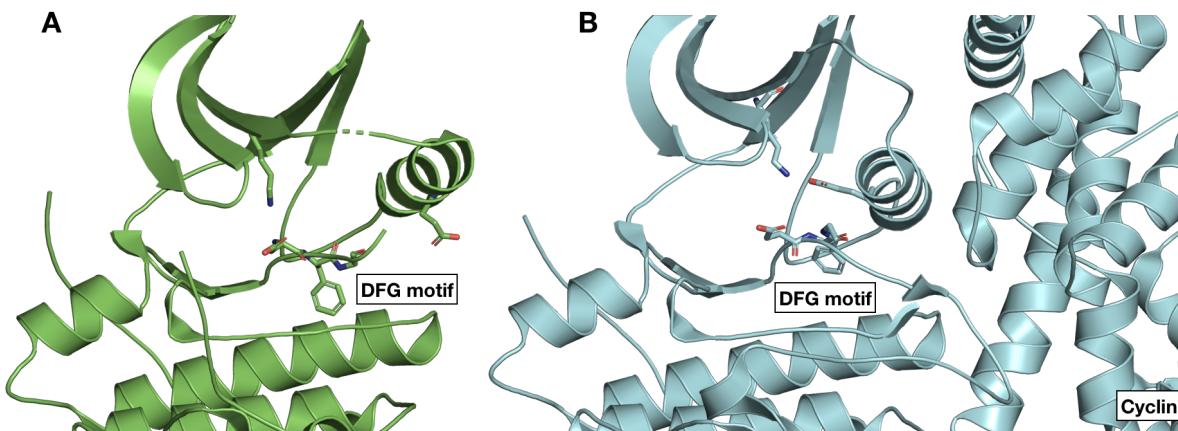
References

- [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. doi: 10.1021/acs.jctc.5b00864.
- [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.

- 467 [20] **Hopkins AL.** Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol.* 2008 Nov; 4(11):682–690.
- 468 [21] **Rudmann DG.** On-target and off-target-based toxicologic effects. *Toxicol Pathol.* 2013 Feb; 41(2):310–314.
- 469 [22] **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.
- 470 [23] **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.
- 471 [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.
- 472 [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.
- 473 [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.
- 474 [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.
- 475 [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.
- 476 [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.
- 477 [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.
- 478 [31] **Din OS,** Woll PJ. Treatment of gastrointestinal stromal tumor: focus on imatinib mesylate. *Ther Clin Risk Manag.* 2008 Feb; 4(1):149–162.
- 479 [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.
- 480 [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.
- 481 [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.
- 482 [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.
- 483 [36] **Robert Roskoski Jr.** USFDA Approved Protein Kinase Inhibitors. . 2017; <http://www.brimr.org/PKI/PKIs.htm>, updated 3 May 2017.
- 484 [37] **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.
- 485 [38] **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.
- 486 [39] **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.
- 487 [40] **Wu P,** Nielsen TE, Clausen MH. FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol Sci.* 2015 Jul; 36(7):422–439.

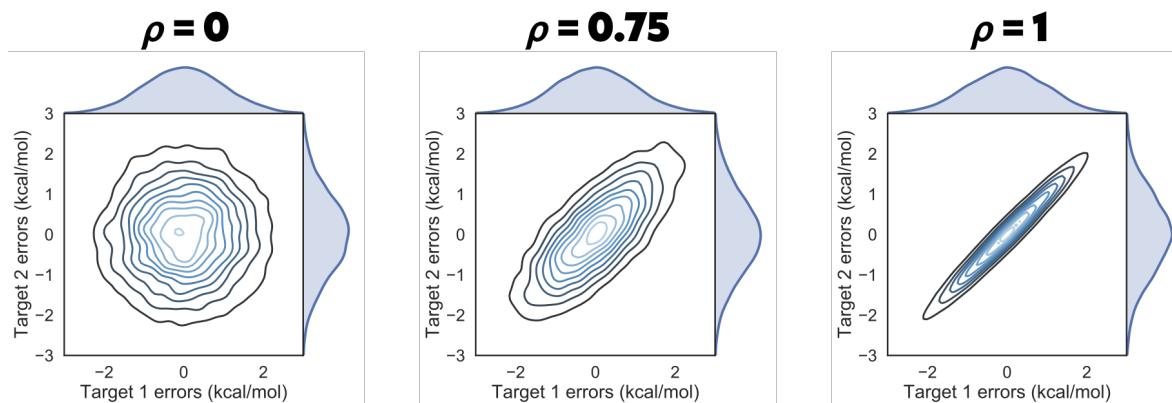
- 513 [41] **Cowan-Jacob SW**, Fendrich G, Floersheimer A, Furet P, Liebetanz J, Rummel G, Rheinberger P, Centeleghe M, Fabbro D,
514 Manley PW, IUCr. Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia.
515 *Acta Crystallogr D Biol Crystallogr.* 2007 Jan; 63(1):80–93.
- 516 [42] **Seeliger MA**, Nagar B, Frank F, Cao X, Henderson MN, Kuriyan J. c-Src Binds to the Cancer Drug Imatinib with an
517 Inactive Abl/c-Kit Conformation and a Distributed Thermodynamic Penalty. *Structure.* 2007 Mar; 15(3):299–311.
- 518 [43] **Huse M**, Kuriyan J. The conformational plasticity of protein kinases. *Cell.* 2002 Jan; 109(3):275–282.
- 519 [44] **Harrison SC**. Variation on an Src-like theme. *Cell.* 2003 Mar; 112(6):737–740.
- 520 [45] **Volkamer A**, Eid S, Turk S, Rippmann F, Fulle S. Identification and Visualization of Kinase-Specific Subpockets. *J Chem
521 Inf Model.* 2016 Feb; 56(2):335–346.
- 522 [46] **Christmann-Franck S**, van Westen GJP, Papadatos G, Beltran Escudie F, Roberts A, Overington JP, Domine D. Un-
523 precedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound-Kinase Activities: A Way
524 toward Selective Promiscuity by Design? *Journal of chemical information and modeling.* 2016 Sep; 56(9):1654–1675.
- 525 [47] **Anastassiadis T**, Deacon SW, Devarajan K, Ma H, Peterson JR. Comprehensive assay of kinase catalytic activity
526 reveals features of kinase inhibitor selectivity. *Nat Biotechnol.* 2011 Nov; 29(11):1039–1045.
- 527 [48] **Davis MI**, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP. Comprehensive
528 Analysis of Kinase Inhibitor Selectivity. *Nat Biotechnol.* 2011 Oct; 29(11):1046–1051. doi: [10.1038/nbt.1990](https://doi.org/10.1038/nbt.1990).
- 529 [49] **Klaeger S**, Heinzlmeir S, Wilhelm M, Polzer H, Vick B, Koenig PA, Reinecke M, Ruprecht B, Petzoldt S, Meng C, Zecha
530 J, Reiter K, Qiao H, Helm D, Koch H, Schoof M, Canevari G, Casale E, Depaolini SR, Feuchtinger A, et al. The target
531 landscape of clinical kinase drugs. *Science.* 2017 Dec; 358(6367).
- 532 [50] **Sun C**, Hobor S, Bertotti A, Zecchin D, Huang S, Galimi F, Cottino F, Prahallas A, Grernrum W, Tzani A, Schlicker A,
533 Wessels LFA, Smit EF, Thunnissen E, Halonen P, Liefink C, Beijersbergen RL, Di Nicolantonio F, Bardelli A, Trusolino L,
534 et al. Intrinsic resistance to MEK inhibition in KRAS mutant lung and colon cancer through transcriptional induction
535 of ERBB3. *Cell Reports.* 2014 Apr; 7(1):86–93.
- 536 [51] **Manchado E**, Weissmueller S, Morris JP, Chen CC, Wullenkord R, Lujambio A, de Stanchina E, Poirier JT, Gainor JF,
537 Corcoran RB, Engelman JA, Rudin CM, Rosen N, Lowe SW. A combinatorial strategy for treating KRAS-mutant lung
538 cancer. *Nature.* 2016 Jun; 534(7609):647–651.
- 539 [52] **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
540 C, Wang S. Substituted 4-(Thiazol-5-yl)-2-(phenylamino)pyrimidines Are Highly Active CDK9 Inhibitors: Synthesis, X-ray
541 Crystal Structures, Structure–Activity Relationship, and Anticancer Activities. *J Med Chem.* 2013 Feb; 56(3):640–659.
- 542 [53] **Blake JF**, Burkard M, Chan J, Chen H, Chou KJ, Diaz D, Dudley DA, Gaudino JJ, Gould SE, Grina J, Hunsaker T, Liu L,
543 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-
544 fluorophenyl)-2-hydroxyethyl)-4-(2-((1-methyl-1H-pyrazol-5-yl)amino)pyrimidin-4-yl)pyridin-2(1H)-one (GDC-0994),
545 an Extracellular Signal-Regulated Kinase 1/2 (ERK1/2) Inhibitor in Early Clinical Development. *J Med Chem.* 2016 Jun;
546 59(12):5650–5660.
- 547 [54] **Hole AJ**, Baumli S, Shao H, Shi S, Huang S, Pepper C, Fischer PM, Wang S, Endicott JA, Noble ME. Comparative Structural
548 and Functional Studies of 4-(Thiazol-5-yl)-2-(phenylamino)pyrimidine-5-carbonitrile CDK9 Inhibitors Suggest the Basis
549 for Isootype Selectivity. *J Med Chem.* 2013 Feb; 56(3):660–670.
- 550 [55] **Berman HM**, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P,
551 Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data
552 Bank. *Acta Crystallogr D Biol Crystallogr.* 2002 Jun; 58(Pt 61):899–907.
- 553 [56] **Sastry GM**, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols,
554 and influence on virtual screening enrichments. *J Comput Aided Mol Des.* 2013 Mar; 27(3):221–234.
- 555 [57] **Harder E**, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL, Kaus JW, Cerutti
556 DS, Krilov G, Jorgensen WL, Abel R, Friesner RA. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small
557 Molecules and Proteins. *J Chem Theory Comput.* 2016 Jan; 12(1):281–296.
- 558 [58] **Abel R**, Wang L, Harder ED, Berne BJ, Friesner RA. Advancing Drug Discovery through Enhanced Free Energy
559 Calculations. *Acc Chem Res.* 2017 Jul; 50(7):1625–1632. doi: [10.1021/acs.accounts.7b00083](https://doi.org/10.1021/acs.accounts.7b00083).

- 560 [59] **Bochevarov AD**, Harder E, Hughes TF, Greenwood JR, Braden DA, Philipp DM, Rinaldo D, Halls MD, Zhang J, Friesner
561 RA. Jaguar: a high-performance quantum chemistry software program with strengths in life and materials sciences.
562 Int J Quantum Chem. 2013; 113(18):2110–2142.
- 563 [60] **Salvatier J**, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Computer Science.
564 2016; 2:e55.
- 565 [61] **Al-Rfou R**, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F, Bayer J, Belikov A, Belopolsky A,
566 Bengio Y, Bergeron A, Bergstra J, Bisson V, Bleecher Snyder J, Bouchard N, Boulanger-Lewandowski N, Bouthillier X,
567 de Brébisson A, Breuleux O, et al. Theano: A Python framework for fast computation of mathematical expressions.
568 arXiv e-prints. 2016 May; abs/1605.02688. <http://arxiv.org/abs/1605.02688>.
- 569 [62] **Bosc N**, Meyer C, Bonnet P. The use of novel selectivity metrics in kinase research. BMC bioinformatics. 2017 Jan;
570 18(1):17.
- 571 [63] **Cheng AC**, Eksterowicz J, Geuns-Meyer S, Sun Y. Analysis of kinase inhibitor selectivity using a thermodynamics-based
572 partition index. J Med Chem. 2010 Jun; 53(11):4502–4510.
- 573 [64] **Hauser K**, Negron C, Albanese SK, Ray S, Steinbrecher T, Abel R, Chodera JD, Wang L. Predicting resistance of clinical
574 Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. Communications Biology. 2018
575 Jun; 1(1):70.
- 576 [65] **Hari SB**, Merritt EA, Maly DJ. Sequence determinants of a specific inactive protein kinase conformation. Chemistry &
577 biology. 2013 Jun; 20(6):806–815.
- 578 [66] **Hu J**, Ahuja LG, Meharena HS, Kannan N, Kornev AP, Taylor SS, Shaw AS. Kinase regulation by hydrophobic spine
579 assembly in cancer. Molecular and cellular biology. 2015 Jan; 35(1):264–276.

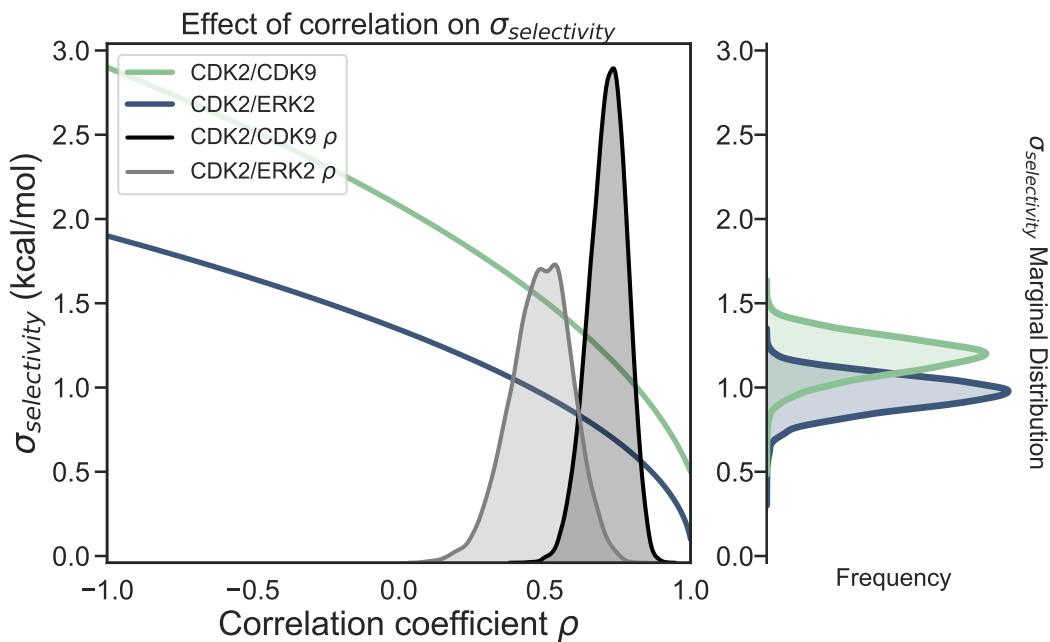
580 **Supplemental Information**

Supplemental Figure 1. 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-out conformation, with the α C 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 α C helix [66] (B) The CDK2 structure used for the CDK2/CDK9 calculations (4BCK) contains cyclin A and adopts a DFG-in/ α C helix-in conformation that forms the salt bridge between K33 and E51. This is typically indicative of a fully active kinase [43, 65].

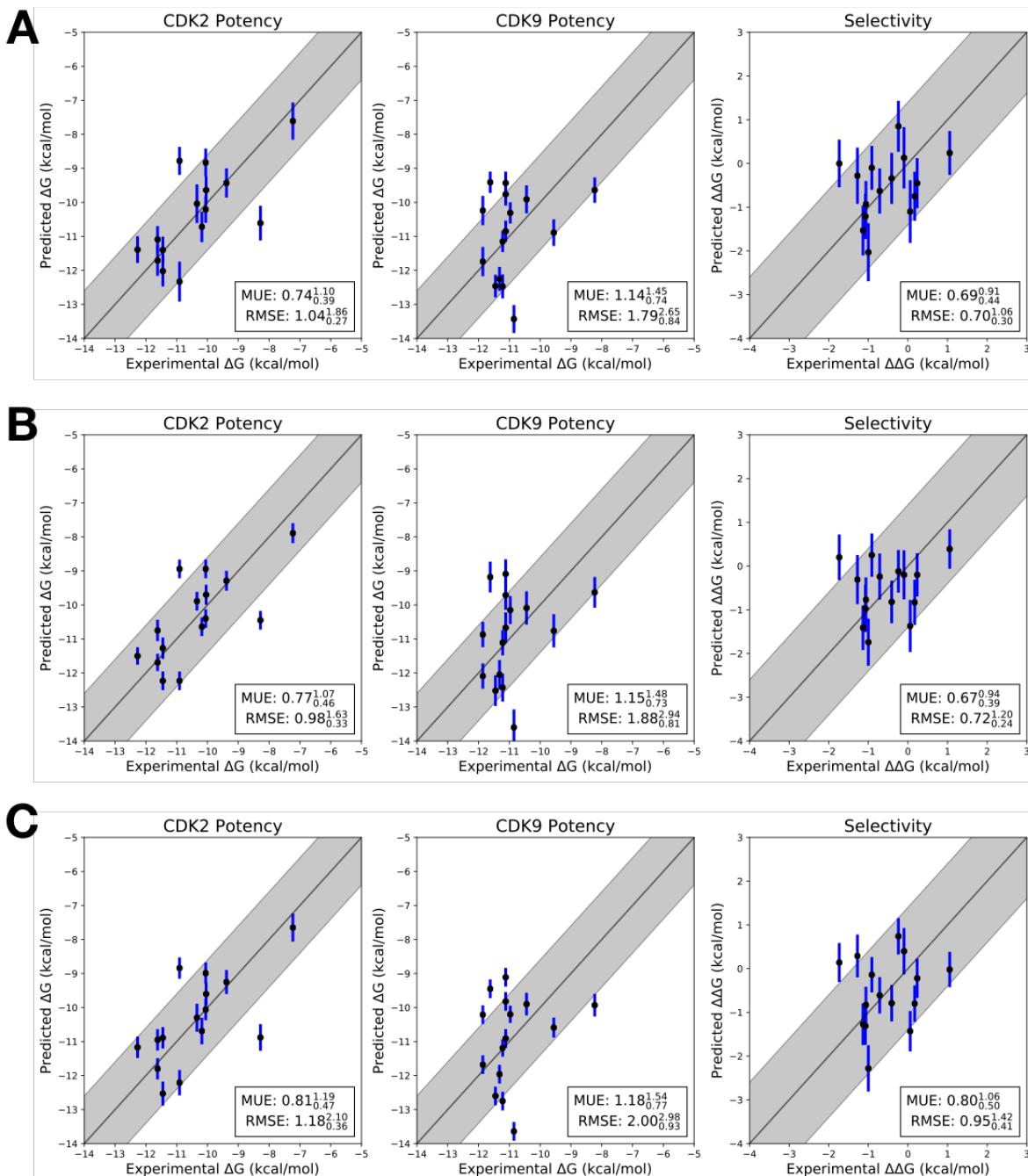


Supplemental Figure 2. Correlation coefficient ρ controls the shape of the joint marginal distribution of errors
As ρ increases, the joint marginal distribution of errors become more diagonal. Each panel shows 10000 samples drawn from a multivariate normal distribution centered around 0 kcal/mol, where the per target error was set to 1 kcal/mol and ρ to the value indicated over the plot.

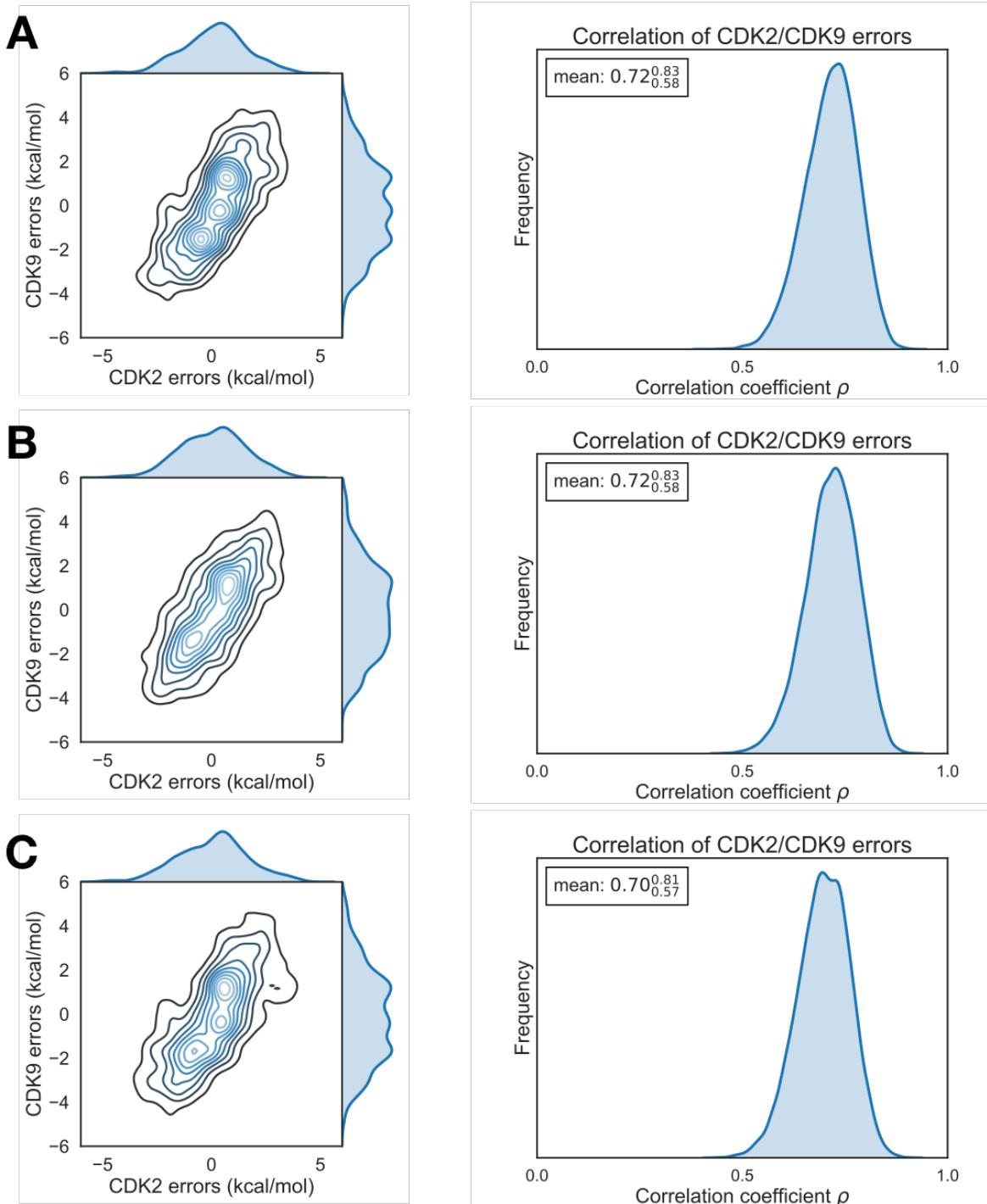


Supplemental Figure 3. Correlation reduces the expected error for selectivity predictions

As corelation coefficient ρ increases, $\sigma_{selectivity}$ decreases. The intersection between CDK2/CDK9 $\sigma_{selectivity}$ (green curve) and ρ (black distribution) indicates the range of expected $\sigma_{selectivity}$ values. The intersection for CDK2/ERK $\sigma_{selectivity}$ (blue curve) and ρ (gray distribution) suggests the expected $\sigma_{selectivity}$ range for that set of calculations. The right side of the plot shows the marginal distribution for CDK2/CDK9 $\sigma_{selectivity}$ (green curve) and CDK2/ERK $\sigma_{selectivity}$ (blue curve).

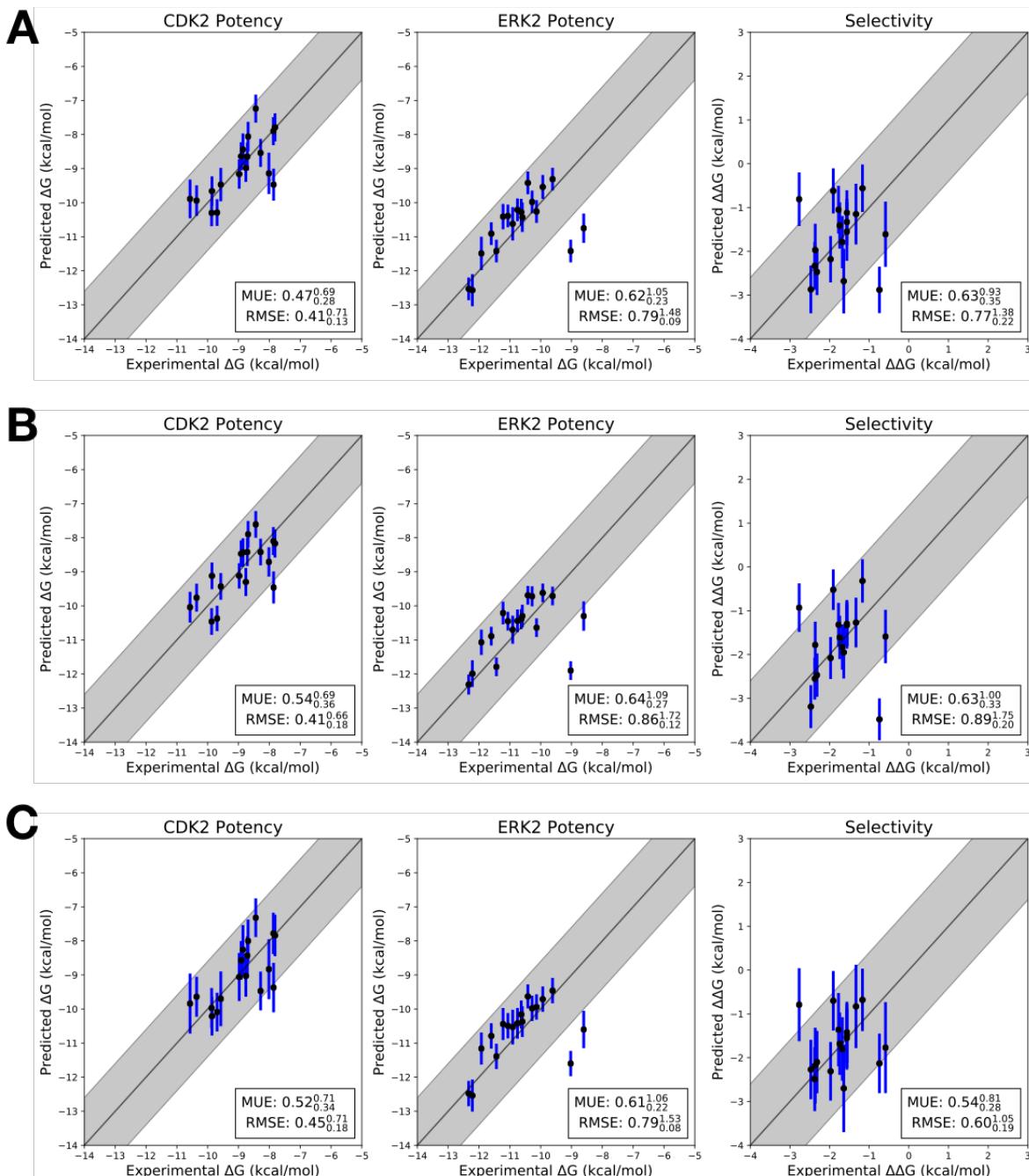
**Supplemental Figure. 4. Each replicate of the CDK2/CDK9 calculations yields a consistent RMSE and MUE**

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 ligand for a given target. All values are shown in units of kcal/mol. The horizontal error bars show the assumed experimental uncertainty of 0.3 kcal/mol[6]. We show the 95% CI based on the estimated statistical as vertical blue error bars. 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

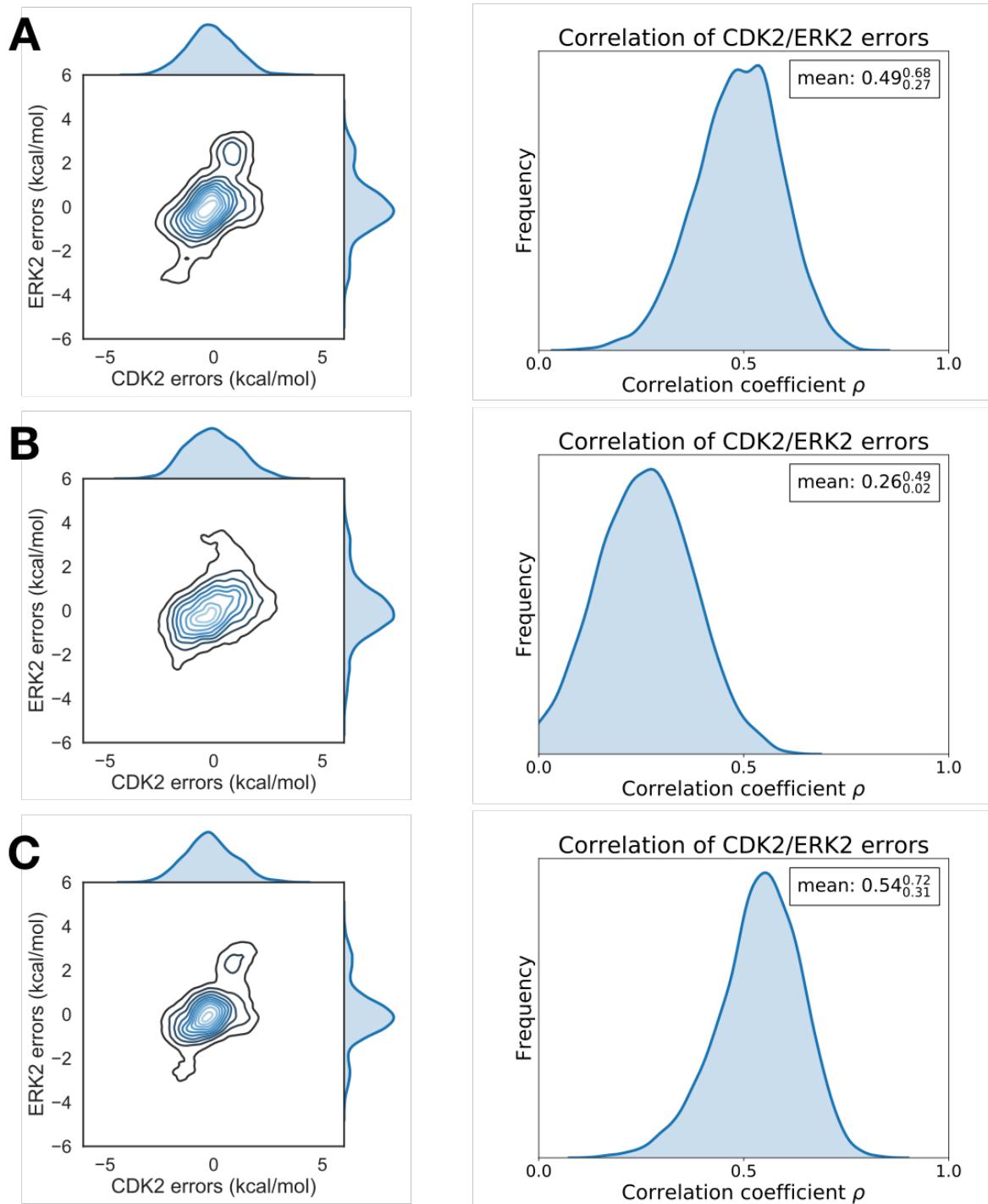


Supplemental Figure 5. Each replicate of the CDK2/CDK9 calculations yields consistent errors and correlation coefficient

(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

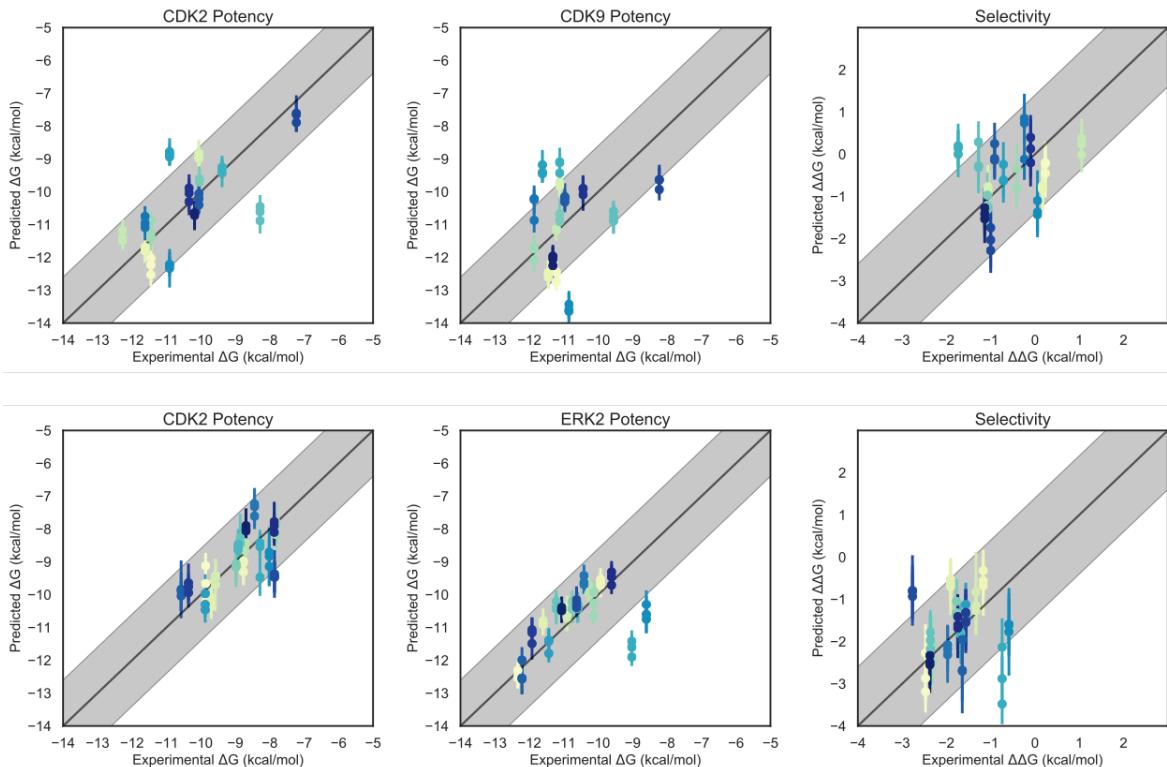
**Supplemental Figure. 6. Each replicate of the CDK2/ERK2 calculations yields a consistent RMSE and MUE**

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 ligand for a given target. All values are shown in units of kcal/mol. The horizontal error bars show the assumed experimental uncertainty of 0.3 kcal/mol [6]. We show the 95% CI based on the estimated statistical as vertical blue error bars. 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



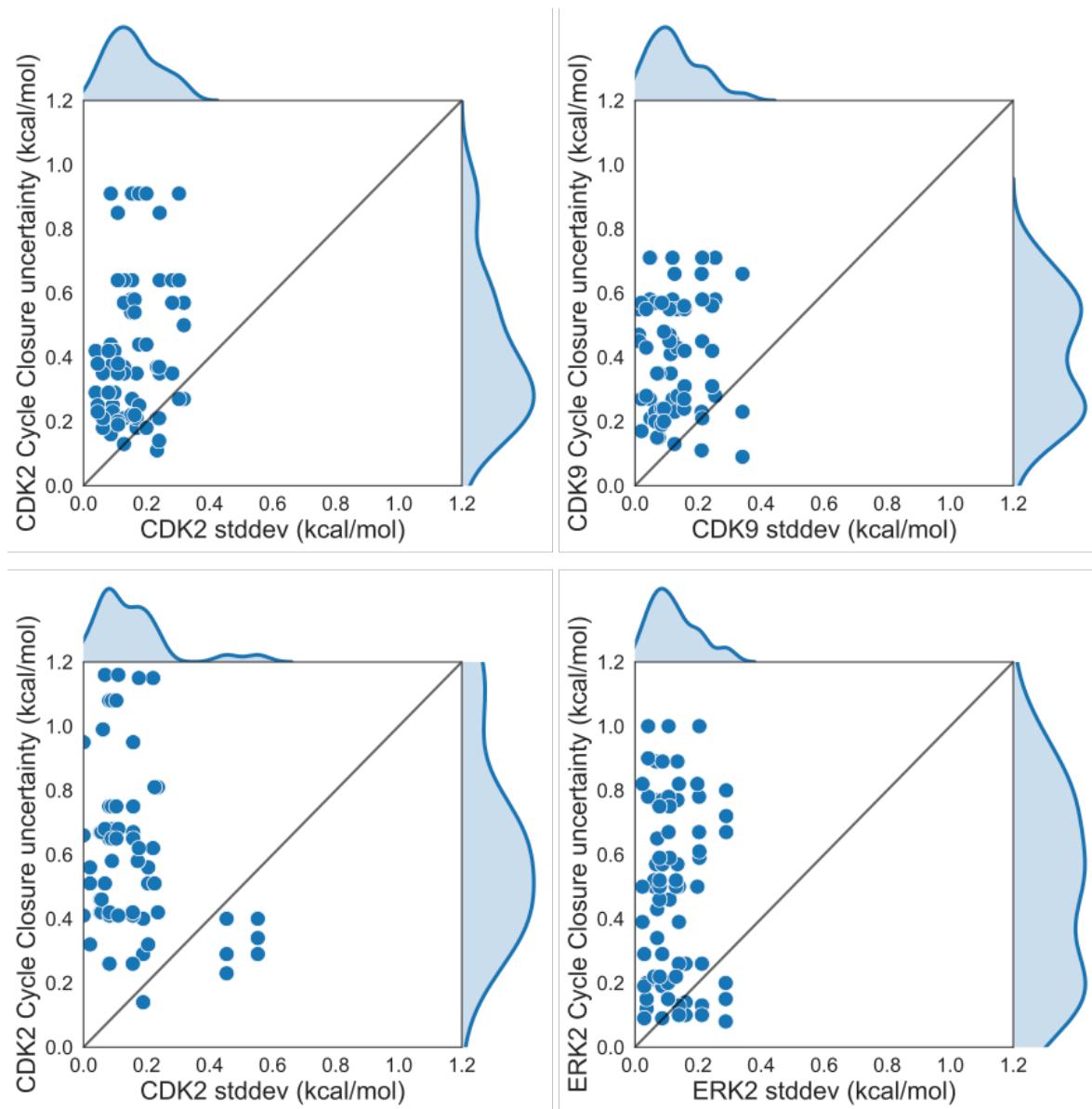
Supplemental Figure 7. Each replicate of the CDK2/ERK2 calculations yields consistent errors and correlation coefficient

(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 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. 8. The pooled replicates show good agreement in predictions for individual ligands

The experimental values are shown on the X-axis and calculated values on the Y-axis. Each data point corresponds to a ligand for a given target. All values are shown in units of kcal/mol. The horizontal error bars show the assumed experimental uncertainty of 0.3 kcal/mol[6]. We show the 95% CI based on the estimated statistical as vertical blue error bars. 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. (**Top**) CDK2/CDK9 replicates (**Bottom**) CDK2/ERK2 replicates



Supplemental Figure. 9. The standard deviation for each edge is smaller than the estimated cycle closure uncertainties

The cycle closure uncertainty for each edge of the map is shown on the Y-axis and the standard deviation for that edge in all three replicate calculations is shown on the X-axis, in kcal/mol. Each point corresponds to an edge of the FEP map. The edges for all three replicates are pooled and shown together. (**Top**) CDK2/CDK9 calculations (**Bottom**) CDK2/ERK2 calculations.