Brooks III Charles L. (Orcid ID: 0000-0002-8149-5417)

QSAR via Multisite λ-Dynamics in the Orphaned TSSK1B Kinase

Xiaorong Liu[1], Pui Ki Tsang[1], Matthew B Soellner[1], and Charles L Brooks III[1, 2, *]

[1]Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, USA

[2]Biophysics Program, University of Michigan, Ann Arbor, Michigan 48109, USA

[*]To whom correspondence should be addressed. E-mail: brookscl@umich.edu Phone: (734) 647-6682

Submitted to *Protein Science* as an *Article*

Abstract

Multisite λ-dynamics (MSλD) is a novel method for the calculation of relative free energies of binding for ligands to their targeted receptors. It can be readily used to examine a large number of molecules with multiple functional groups at multiple sites around a common core. This makes MSλD a powerful tool in structure-based drug design. In the present study, MSλD is applied to calculate the relative binding free energies of 1296 inhibitors to the testis specific serine kinase 1B (TSSK1B), a validated target for male contraception. For this system, MSλD requires significantly fewer computational resources compared to traditional free energy methods like free energy perturbation or thermodynamic integration. From MSλD simulations, we examined whether modifications of a ligand at two different sites are coupled or not. Based on our calculations, we established a quantitative structure-activity relationship (QSAR) for this set of molecules and identified a site in the ligand where further modification, such as adding more polar groups, may lead to increased binding affinity.

Introduction

Alchemical free energy calculations are powerful tools in computer-aided drug design, especially in the lead optimization process. For example, free energy perturbation (FEP) [1, 2] and thermodynamic integration (TI) [3] are among the most popular alchemical free energy methods in structure-based drug design. Instead of directly computing absolute binding free energies for two ligands, where sufficient sampling is often difficult because ligand binding is a slow and complex process, the relative binding free energies can be obtained by considering alchemical transformations between two ligands in both an unbound state and a protein bound state (**Figure 1**). In combination with recent advancements in developing more accurate force fields (and powerful machine learning strategies in some cases), these alchemical free energy calculations have gained a lot of success in accelerating drug development [4-9]. Depending on the system of interest, the errors of calculated relative binding free energies are usually less than 1.5 kcal/mol, which implies that these approaches are practically useful and can offer significant advantage in the lead optimization workflow [10].

$$\begin{tikzcd}
P + L1 \arrow[r, "\Delta G_{bind}(L1)"] \arrow[d, "\Delta G_{unbound}(L1\rightarrow L2)"'] & P\bullet L1 \arrow[d, "\Delta G_{bound}(L1\rightarrow L2)"] \\
P + L2 \arrow[r, "\Delta G_{bind}(L2)"'] & P\bullet L2
\end{tikzcd}$$

$$\begin{aligned}
&\Delta\Delta G_{bind}(L1\rightarrow L2) \\
&= \Delta G_{bind}(L2) - \Delta G_{bind}(L1) \\
&= \Delta G_{bound}(L1\rightarrow L2) - \Delta G_{unbound}(L1\rightarrow L2)
\end{aligned}$$

Figure 1. *Thermodynamic cycle for calculating relative binding free energy between two ligands (L1 and L2) to a common protein (P). Alchemical methods consider the two vertical processes rather than the two horizontal processes, and it is particularly useful for two congeneric ligands.*

Although highly accurate, a major limitation of TI and FEP is the computational cost. In both TI and FEP, the alchemical transformation between two ligands is divided into multiple steps (or windows) to ensure sufficient overlap of phase space between two adjacent windows. Usually, 10-20 λ windows are needed in one set of TI or FEP simulations and more λ windows are needed if there is a change in the net charge of two ligands, which makes them expensive [11]. Moreover, these methods scale poorly with the total number of compounds studied. To compute relative binding free energies between *N* molecules, at least *N*-1 sets of FEP or TI simulations are needed and typically more are used to ensure cycle closure using methods like LOMAP[12]. For this reason, it is very expensive to compute relative binding free energies for hundreds or thousands of ligands to a common protein target using these approaches.

Multisite λ dynamics (MSλD) is an innovative alchemical free energy method that aims, in part, to address these two problems [13, 14]. In MSλD, the alchemical coupling parameter λ is treated as a dynamic variable that can spontaneously fluctuate between two end states (i.e., $\lambda = 0$ and $\lambda = 1$). In this way, we do not need multiple λ windows to compute the free energy difference between two molecules. Moreover, multiple λ parameters corresponding to multiple functional groups at different sites around a common core can be used. This enables us to explore a combinatorial chemical space of hundreds to thousands of small molecules in a single MSλD simulation. In this way, MSλD addresses the scalability problem encountered in FEP and TI calculations, thus significantly reducing the computational cost of accurate free energy calculations. Moreover, it has been demonstrated that MSλD does so without any loss in statistical precision [9, 14-16].

In the present study, we apply MSλD to calculate relative binding free energies of 1296 inhibitors to testis specific serine kinase 1B (TSSK1B), a validated target for male contraception. TSSK1B is primarily expressed in the testis and is required during spermatid development. Overexpression of TSSK1B has been reported in human cancers, including some colon, bladder, and breast cancers [17]. Its missense mutations have been discovered in male patients with unresolved infertility causes and are associated with sperm dysmorphology [18]. However, TSSK1B has been identified as one of the “understudied” kinases by Illuminating the Druggable Genome (IDG), since it lacks antibodies and validated chemical probes [19]. An initial screening assay using the published kinase inhibitor set (PKIS)[20] identified that the molecule KQQ (see **Figure 2**) inhibits

TSSK1B activity with a cellular half maximal inhibitory concentration (IC_{50}) of 180 nM, which suggests that molecule KQQ is a lead compound for targeting TSSK1B.

To find KQQ analogs that have improved potency for TSSK1B, we would like to perform large scale free energy calculations to study the quantitative structure-activity relationship (QSAR). As an initial attempt to achieve this goal, the size and shape of the substituent at sites A, C, and D of the ligand was varied from hydrogen through methyl, ethyl, butyl, isopropyl, to tert-butyl, as illustrated in **Figure 2**. This allowed us to probe the size and shape of the binding pocket in TSSK1B and identify potential hydrophobic contacts. Site B could be either carbon or nitrogen, further extending the explored chemical space. Three substituents were tested at site E, since the relevant experimental measurements are available. Taken together, we examined the relative binding free energies of 6*2*6*6*3=1296 compounds using MSλD. To the best of our knowledge, this may represent one of the largest scale free energy calculations to date. Thus, the current study demonstrates the unique advantages of the MSλD free energy framework, which enables us to rapidly explore this vast chemical space and obtain QSAR of a combinatorial library of small molecules.

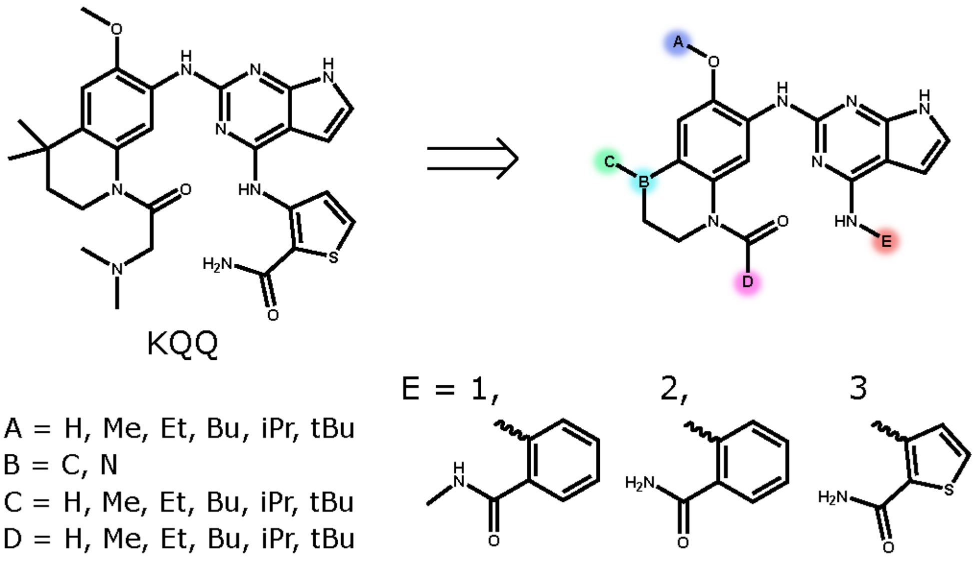
KQQ
A
B
C
D
E
A = H, Me, Et, Bu, iPr, tBu
B = C, N
C = H, Me, Et, Bu, iPr, tBu
D = H, Me, Et, Bu, iPr, tBu
E = 1,
2,
3

Figure 2. Structure of lead compound KQQ and its modifications. There are six substituents at sites A, C and D, two substituents at site B, and three substituents at site E. This gives a combinatorial library of 1296 unique molecules.

Methods

Homology modeling of TSSK1B structure

As an understudied kinase, there is no experimental structure of TSSK1B to date. Therefore, we used homology modeling to construct a model for the TSSK1B structure. Four modeling methods were tested, Swiss Model [21], I-TASSER [22-24], Robetta [25] and AlphaFold 2 [26], and all predicted models were highly similar, containing a well-structured region (residues 1-275) followed by a long, disordered tail (residues 276-367) (**Figure S1**). The root mean squared deviation (RMSD) of $C\alpha$ atoms in folded regions between any two homology models ranged from 1.8 to 2.1 Å (i.e., excluding loop residues 150-180 and 276-367). Moreover, the structure of residues near the binding site are highly similar (**Figure S2**). Therefore, we assume that the specific choice of homology modeling method will have little impact on the resultant TSSK1B structure and subsequent free energy calculations. Since the C-terminal tail (residues 276-367) is highly disordered, it poses great challenges in accurate modeling of its structural ensemble as well as sufficient sampling of its heterogeneous conformations in molecular dynamics (MD) simulations, so this region was discarded in our MSλD simulations. This is expected to introduce little artifact in relative binding free energy calculations since MD simulations of the whole protein (residues 1-367) in complex with the lead compound KQQ suggest that this disordered C-terminal tail does not form stable contacts with the ligand (**Figure S3**).

System preparation

In MSλD simulations, we construct a hybrid ligand where all atoms that belong to the maximum common substructure are represented once and atoms unique to each ligand are represented explicitly [14]. We have used the automated workflow named *msld_py_prep* (https://github.com/Vilseck-Lab/msld-py-prep) to identify the maximum common substructure in each simulation, and partial atomic charges in the maximum common substructure were assigned using the charge renormalization scheme [27]. Initial conformations of these small molecules were generated by modifying the ligand structure extracted from PDB 4PNI [28], an experimental structure of G protein-coupled receptor kinase 1 in complex with the lead compound KQQ. The CHARMM general force field (CGenFF) [29, 30] was used to model these compounds. In all MD simulations, the initial structure of TSSK1B (residues 8-275) was obtained from Swiss Model [21] using 4PNI as the template, with the N- and C-termini of the truncated construct (residues 8-275 only) capped with an N-acetyl group and N-methyl amide, respectively. Protonation states of titratable residues were determined using PROPKA3 [31, 32]. This analysis suggested that both Nδ and Nε of His109 were protonated at pH 7.4, the pH representing our simulated solution. For simulations of TSSK1B in complex with small molecules, the ligands were initially placed into the TSSK1B binding pocket based on the binding pose observed in PDB 4PNI. The protein was modeled using the CHARMM36m force field [33]. Each system was solvated in a cubic box using TIP3P [34] water molecules, allowing for at least 12 Å between protein/ligand and the nearest box edge. To neutralize the system and represent a NaCl concentration close to the physiological concentration (i.e., 150 mM), a proper amount of NA^+ and Cl^- ions were randomly placed in bulk water.

MSλD methodology

The details of MSλD methodology have been reported in previously published works [14, 35]. In brief, the potential energy of the system is as follows:

$$V = V(x_0, x_0) + \sum_{s=1}^{M}\sum_{i=1}^{N_s}\lambda_{si}(V(x_0, x_{si}) + V(x_{si}, x_{si})) + \sum_{s=1}^{M}\sum_{i=1}^{N_s}\sum_{t=s+1}^{M}\sum_{j=1}^{N_t}\lambda_{si}\lambda_{tj}V\left(x_{si}, x_{tj}\right) + V_{bias}(\{\lambda\}) \qquad (1)$$

where x_0 are the coordinates of the environment, including water molecules, ions, the maximum common substructure of the hybrid ligand and the protein, if present. x_{si} are ligand coordinates of substituent *i* at site *s*. *M* is the total number of sites, and N_s is the number of substituents at site *s*. V_{bias} is the biasing potential used to facilitate transitions between different end states.

To maximally enhance sampling in alchemical space, optimized biasing potentials, V_{bias}, need to be used in MSλD simulations as described in previous work [35, 36]. Briefly, four types of biasing potentials are applied, V_{fixed}, V_{quad}, V_{end} and V_{skew}, as shown below. V_{fixed} is to ensure that the end points have similar free energies. V_{quad} is tuned to lower the free energy barrier between two end points. V_{end} and V_{skew} are used to help escape the deep free energy wells near the end point.

$$V_{bias} = V_{fixed} + V_{quad} + V_{end} + V_{skew} \qquad (2)$$

$$V_{fixed} = \sum_{s}^{M} \sum_{i}^{N_s} \phi_{si}\lambda_{si} \qquad (3)$$

$$V_{quad} = \sum_{s}^{M} \sum_{i}^{N_s} \sum_{j>i}^{N_s} \psi_{si,sj}\lambda_{si}\lambda_{sj} \qquad (4)$$

$$V_{end} = \sum_{s}^{M} \sum_{i}^{N_s} \sum_{j\neq i}^{N_s} \omega_{si,sj}\lambda_{si}\lambda_{sj}/(\lambda_{si} + \alpha) \qquad (5)$$

$$V_{skew} = \sum_{s}^{M} \sum_{i}^{N_s} \sum_{j\neq i}^{N_s} \chi_{si,sj}\lambda_{si}(1 - exp(-\lambda_{si}/\delta)) \qquad (6)$$

where $\alpha = 0.017$ and $\delta = 0.18$. Other parameters, $\{\phi_{si}\}$, $\{\psi_{si,sj}\}$, $\{\omega_{si,sj}\}$and $\{\chi_{si,sj}\}$ are optimized using adaptive landscape flattening (ALF, available at https://github.com/RyanLeeHayes/ALF) [35, 36], an iterative process to flatten free energy landscapes in the alchemical space. At each iteration, the system was equilibrated, followed by short MD simulations. Two phases of biasing potential optimization were used in this work. In the first phase, 50 iterations of 100-ps simulations were performed, and in the second phase, 14 or more iterations of 1-ns simulations were carried out until reversible transitions can be observed in alchemical space. Sampling statistics were used to compute three types of free energy profiles, including $G(\lambda_{si})$, $G(\lambda_{si}, \lambda_{sj})$, and $G(\lambda_{si} /(\lambda_{si} + \lambda_{sj}))$ under the condition of $\lambda_{si} + \lambda_{sj} > 0.8$. Parameters $\{\phi_{si}\}$, $\{\psi_{si,sj}\}$, $\{\omega_{si,sj}\}$and $\{\chi_{si,sj}\}$ were updated to flatten these three types of free energy profiles. Note that implicit constraints in **Eq. 7** bias sampling towards the end points [37], which is a desirable property for efficient calculation of free energy differences between two ligands based on the empirical estimator in **Eq. 8**. Therefore, the corresponding entropic component of free energies originating from these implicit constraints was removed from total free energies before flatting the free energy profiles.

$$\lambda_{si} = \frac{exp(5.5sin\theta_{si})}{\sum_{j=1}^{N_s} exp(5.5sin\theta_{sj})} \qquad (7)$$

$$\Delta G^{MS\lambda D}(\{\lambda_{si}\} \rightarrow \{\lambda_{sj}\}) \approx -k_B T ln \frac{P(\{\lambda_{sj}\} > 0.99)}{P(\{\lambda_{si}\} > 0.99)} - [V_{bias}(\{\lambda_{sj}\} = 1) - V_{bias}(\{\lambda_{si}\} = 1)]$$

(8)

Simulation details

All MD simulations were performed using CHARMM v46a1 [38, 39]. The DOMDEC package [40] within CHARMM was utilized to accelerate simulations on GPUs. Periodic boundary conditions were applied to all systems. The van der Waals interactions were smoothly switched to zero between 9 and 10 Å, using the VFSWITCH function in CHARMM. Electrostatic interactions were treated with the particle mesh Ewald method [41]. To alleviate the problem of solvent inaccessible cavities as λ approaches 0 when hard-core potentials are used, a soft-core potential [35] was used in the MSλD simulations. All simulations were performed under NPT conditions, with a temperature of 298.15 K and a pressure of 1 atm. Any bond containing hydrogen atoms was constrained using SHAKE [42], allowing for an integration time step of 2 fs. Five independent MSλD simulations were run for each system (see **Table 1**). Since the free ligand simulations converged more rapidly than the simulations of the protein-ligand complex, the production runs of the former were shorter than those of the latter. Examination of λ values as a function of simulation time suggests that sufficient sampling was achieved in alchemical space (**Figure S4**). Free energy uncertainties were estimated using bootstrap analysis. Specifically, for each of the five independent simulations, free energies were estimated using the λ trajectories. 50 free energy estimates were obtained from these five independent measurements using random sampling with replacement, and standard deviation of these 50 estimates was taken to be the uncertainty.

Table 1. *MSλD simulations performed in this work. Note that for each set, we have carried out simulations in both unbound and protein bound state as illustrated in* ***Figure 1****.*

Set	Simulation Time for each Replicate (ns)		Hybrid Ligand				
	unbound	bound	site A	site B	site C	site D	site E
I	10	40	same as KQQ	same as KQQ	same as KQQ	same as KQQ	1, 2, 3
II*	40	120	H, Me, Et, Bu, iPr, tBu	C	H, Me, Et, Bu, iPr, tBu	H, Me, Et, Bu, iPr, tBu	1, 2, 3
III*	40	120	H, Me, Et, Bu, iPr, tBu	N	H, Me, Et, Bu, iPr, tBu	H, Me, Et, Bu, iPr, tBu	1, 2, 3
				C	H		
IV	10	40	same as KQQ	same as KQQ	same as KQQ	same as KQQ	Modified substituent 1

*Since sites B and C are directly connected (see **Figure 2**) and probably tightly coupled, we can consider site B-C as one site with 2*6=12 substituents. In principle, we can perform a single simulation with all 12 substituents at site B-C. However, the fraction of physical ligand decreases if there are too many substituents at a given site. Therefore, here we performed two simulations to alleviate this problem and to reduce computational cost. There is no fixed rule about how to divide these 12 substituents among multiple simulations, as long as there is at least one common ligand between them so that we can combine the simulation results to compute relative binding free energies of all possible ligands.

Experimental measurements of IC_{50}

We have measured IC_{50} for three ligands where sites A, B, C and D are the same as those in molecule KQQ and site E is one of the three substituents in **Figure 2**. The TSSK1B NanoBRET TE assay kit (NV4471) was purchased from Promega and carried out as described in the assay kit. HEK293 cells (ATCC) were used for transfection. Commercial nanoBRET tracer K9 (N2632) was used for the tracer at 660 nM. The adherent cell format was used as this led to optimal BRET. BRET ratios were calculated from the donor signal (415 nm) and acceptor signal (610 nm).

Results and discussion

Comparison of simulation results with experimental measurements

IC_{50} values for three of the ligands in which sites A, B, C and D are the same as those in molecule KQQ and site E is one of the three substituents shown in **Figure 2**, have been measured experimentally (see **Figure S5**). This allows us to examine the force field accuracy in reproducing their relative binding affinities. As shown in **Table 2**, our MSλD simulations (set I in **Table 1**) suggest that substituent 3 at site E leads to strongest binding, while binding becomes weaker with substituent 1 or 2. But experimentally, the tightest binder to TSSK1B is when site E is substituent 2, followed by those with substituent 3 and 1, whose binding free energies are 0.4 and 2.2 kcal/mol higher than that with substituent 2, respectively. Such discrepancy was partly due to the imperfect dihedral potentials for rotatable bonds that connect substituents at site E with the maximum common substructure. As shown in **Table 2**, after we optimize dihedral potentials based on quantum mechanics calculations (see **Figure S6**), our MSλD calculations better captured the overall trend that substituents 2 and 3 at site E result in tighter binding than substituent 1, although the simulation results still underestimate the improvement of binding affinity when replacing substituent 1 with the other two. Note that the experimental measurements of IC_{50} were conducted using the nanoBRET assay with intact cells, where any difference in permeability across the cell membrane between these molecules could also contribute to the difference in IC_{50} values. Also, compound KQQ is a type-I inhibitor, which occupies the ATP binding site. In the nanoBRET assay, ATP competitively binds TSSK1B and this will have an impact on the IC_{50} measurement. It is also possible that we did not obtain sufficient sampling of certain important structural reorganizations of protein and/or water molecules. For instance, based on the PDB structure 4PNI, there are two water molecules inside the binding pocket that are within 5 Å of site E of molecule KQQ. In our simulations, although we also observed multiple water molecules near site E, it is possible that the probability of water molecules residing inside the active site is different for different substituents at site E, and insufficient sampling of motion of these water molecules may contribute to the discrepancy between simulation and experimental results. Moreover, errors of order of 1 kcal/mol for free energy calculations are near the anticipated force field limit. For instance, previous MSλD simulations of β-secretase 1 inhibitors using the CGenFF force field in combination with CM1A partial atomic charges illustrated that simulation results deviate from experimental measurements by a mean unsigned error (MUE) of 0.47 kcal/mol [15]. A recent study of a large number of ligands for seven proteins demonstrated that the MUE may range from 0.39 to 0.93 kcal/mol, depending on the protein of interest [9]. A non-equilibrium thermodynamic integration method with Amber and CHARMM force fields has also been used to examine a large dataset of 482 ligand modifications from 13 different protein-ligand datasets, and the overall MUE was found to be 0.87 kcal/mol, with MUE for each protein-ligand system ranging from 0.47 to 1.26 kcal/mol [43]. Since our ultimate goal is to find KQQ analogs that show significantly higher binding affinities to TSSK1B, we did not further optimize CGenFF to improve its accuracy, and the above optimized dihedral potentials (see **Figure S6**) were used in all later simulations.

Table 2. *Relative binding free energies for modification of KQQ molecule at site E. The molecule in which site E is substituent 1 is selected as the reference molecule for computing relative binding free energies.* ΔG_{bind}^{exp} *was estimated based on RTln(IC*$_{50}$*).*

Substituent Index	1	2	3
IC$_{50}$ (nM)	3800 ± 1900	90 ± 40	180 ± 90
$\Delta\Delta G_{bind}^{exp}$ (kcal/mol)	0	-2.23	-1.82
$\Delta\Delta G_{bind}^{MS\lambda D}$ (kcal/mol)	0 ± 0.10	-0.20 ± 0.09	-1.04 ± 0.21
$\Delta\Delta G_{bind}^{MS\lambda D}*$ (kcal/mol)	0 ± 0.28	-0.38 ± 0.19	-0.47 ± 0.25

* MSλD simulation using CGenFF with optimized dihedral potentials. To see if the relative binding free energies are significantly different when site E is substituent 1 or 2, we performed independent samples t-test. The t value was calculated from $(\bar{X}_1 - \bar{X}_2)/(\sqrt{(s_1^2 + s_1^2)/n})$, where $\bar{X}_1$ and $\bar{X}_2$ are mean values, s_1 and s_2 are standard deviations, and n is 5. The calculated t value was 2.511, which was greater than the two-tailed table value of 2.306 at $\alpha = 0.05$, suggesting that the binding free energies can be considered different between substituents 1 and 2. Similarly, the difference in binding free energy between substituents 1 and 3 can also be considered statistically significant since the calculated t value is 2.800, greater than table value of 2.306.

Sites A, C, D and E are largely independent of each other

A commonly used assumption in many small molecule drug design projects is that two sites are independent of each other, and therefore effects of modifications on different sites are additive. To test whether this is a reasonable assumption in the present study, we compared the calculated relative binding free energies obtained from this additive model (**Eq. 9**) with those directly from MSλD simulations (**Eq. 8**) for all combinations of two sites. As demonstrated in **Figure 3**, relative binding free energies obtained from these two methods are in good agreement with each other, suggesting that sites A, C, D and E are not strongly coupled, and we can use the additive model to estimate relative binding free energies.

$$\Delta G^{additive}\left(\{\lambda_{si}\} \rightarrow \{\lambda_{sj}\}\right) \approx \sum_{s=1}^{M}\{-k_B T ln \frac{P(\lambda_{sj}>0.99)}{P(\lambda_{si}>0.99)} - [V_{bias}(\lambda_{sj} = 1) - V_{bias}(\lambda_{si} = 1)]]\} \quad (9)$$

Note that we do not need to make any assumption that two sites are independent *before* setting up a MSλD simulation. After obtaining a MSλD simulation trajectory, we can always check if two sites are independent like in **Figure 3**. If site *i* and site *j* are coupled, we could use Potts model-based estimator to compute free energies [44]. Alternatively, we may treat sites *i* and *j* as a single site with N_i*N_j substituents and test if this new site is independent of other sites. This process can be repeated if more than two sites are coupled with each other. One advantage of this approach is to further reduce sampling requirements of MSλD simulations. Specifically, since the probability of observing $\lambda_{si} > 0.99$ and $\lambda_{sj} > 0.99$ *simultaneously* is always lower than the probability of observing $\lambda_{si} > 0.99$, the uncertainty associated with $\Delta G^{MS\lambda D}$ in **Eq. 8** is greater than the uncertainties of $\Delta G^{additive}$ in **Eq. 9** for a given MSλD simulation. As demonstrated in previous work [44], if there is no coupling between any two sites, ALF is able to completely flatten the alchemical free energy landscape, and all sites have the same relaxation time scale, the amount of sampling for the original estimator (**Eq. 8**) scales exponentially to the number of sites (M). For Potts model including up to two-body terms, the sampling requirement scales to M^2. For the additive model (**Eq. 9**), the sampling should be proportional to M. Therefore, to achieve certain precision, the additive model presents a more rapid approach of using MSλD to calculate relative free energies whenever possible. In this way, MSλD simulation in combination with the additive model could significantly improve the computational efficiency of free energy calculations when exploring large combinatorial chemical space.

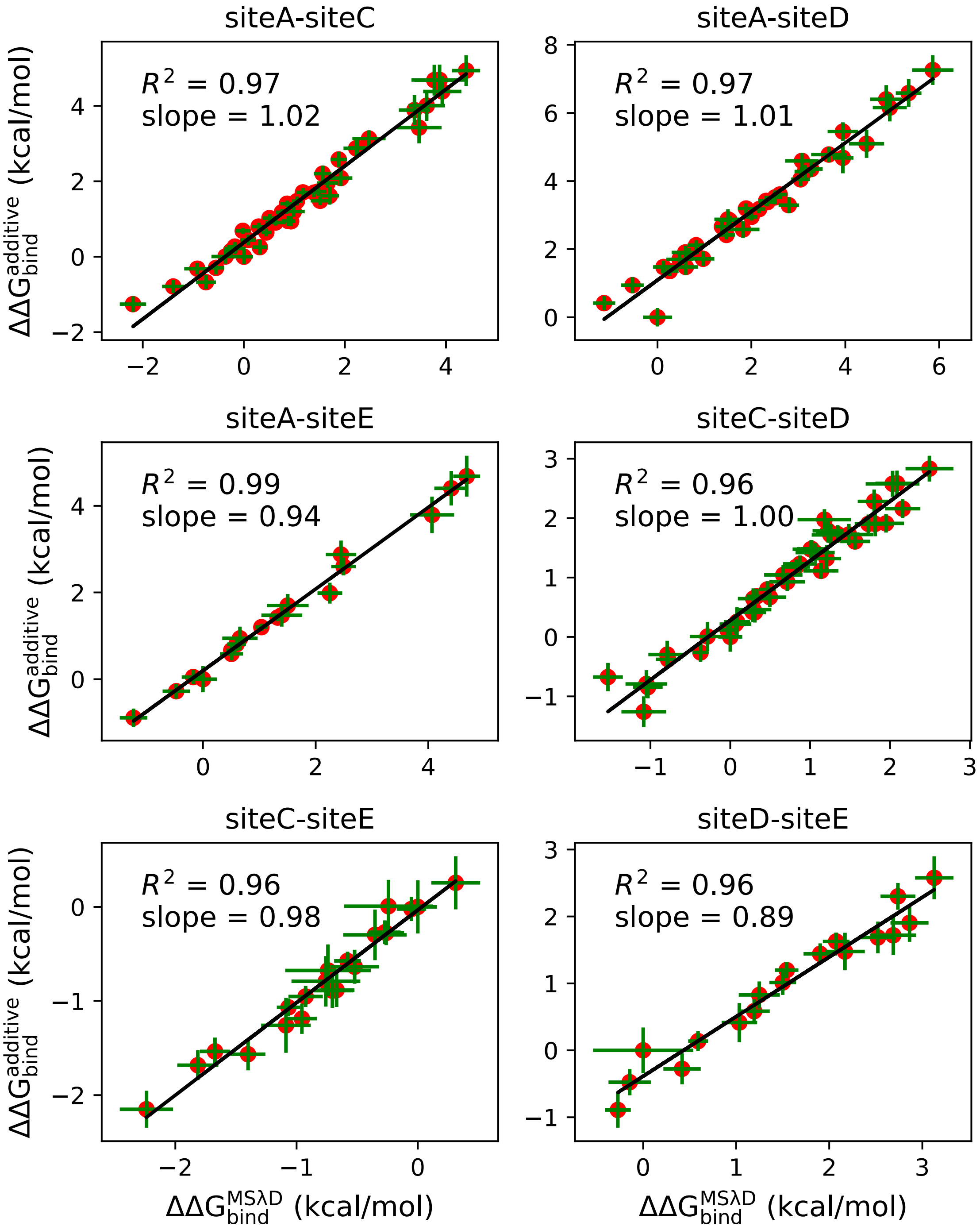

Figure 3. *Comparison of relative binding free energies computed from the standard estimator in* ***Eq. 8*** *and the additive model in* ***Eq. 9****. Results were computed from simulation set III. See* ***Figure S7*** *for results from simulation set II.*

Structure-activity relationship for KQQ variants binding to TSSK1B

The above findings that site A, C, D and E are independent of each other greatly simplified our QSAR study and further design. Now we can examine each site individually and determine which site(s) we should focus on and what additional modifications may be needed based on the structure-activity relationships. As shown in **Figure 4**, whether site B is carbon or nitrogen, varying the substituent at site A between methyl, ethyl, butyl, isopropyl and tert-butyl shows the same trend, where tert-butyl appears to be most unfavorable and the butyl group is most favorable among these hydrophobic substituents. To understand the structural basis of these free energy results, we examined how site A substituents fit and interact with the TSSK1B binding pocket. As shown in **Figures 4**, protein residues near site A are mainly hydrophobic and this pocket is largely solvent exposed. Small/linear alkyl groups nicely fit in the hydrophobic groove (**Figure 4C** for example). Branched alkyl groups are not able to fit into the protein pocket very well due to steric hindrance, thus leading to reduced binding affinities.

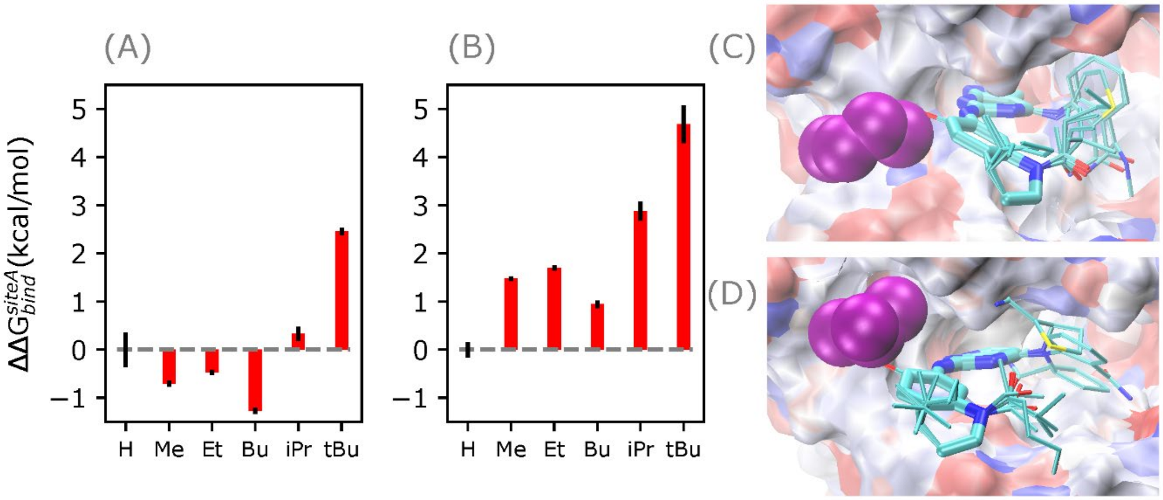

__Figure 4.__ Relative binding free energies when substituent at site A is varied, while site B is either carbon (A) or nitrogen (B). The results were obtained from simulation set II and III, respectively, and can be found in __Table S1__ as well. (C) Illustration of how butyl at site A fits the TSSK1B binding pocket. (D) Illustration of how tert-butyl at site A fits the TSSK1B binding pocket. Here, the protein is shown as a surface, colored by partial atomic charges, with red for negative charges and blue for positive charges. Ligand is shown in Licorice, with substituent at site A highlighted in purple van der Waals spheres.

We also found that changing substituent at site C between hydrogen, methyl, ethyl, butyl, isopropyl and tert-butyl groups has little impact on binding free energies, regardless of whether site B is carbon or nitrogen (see **Figure 5**). Again, the main reason is that site C is mostly solvent exposed (see **Figure 5C-D** for example). The general trend of free energy among these substituents appears to be slightly different when site B is changed from carbon to nitrogen. Further examination of simulation trajectories suggests that when site B is carbon, the site C substituent, such as the butyl group, mainly pointed "up" during simulations (**Figure 5C**). Having a branched alkyl group, like isopropyl or tert-butyl, at site C may lead to steric hindrance, thus reducing binding affinity. In contrast, when site B is nitrogen, the tert-butyl group, for example, mainly pointed "down" in the simulation. In this way, the site C substituent is mostly solvent exposed and there is more room to accommodate a bulky group like isopropyl or tert-butyl (**Figure 5D**).

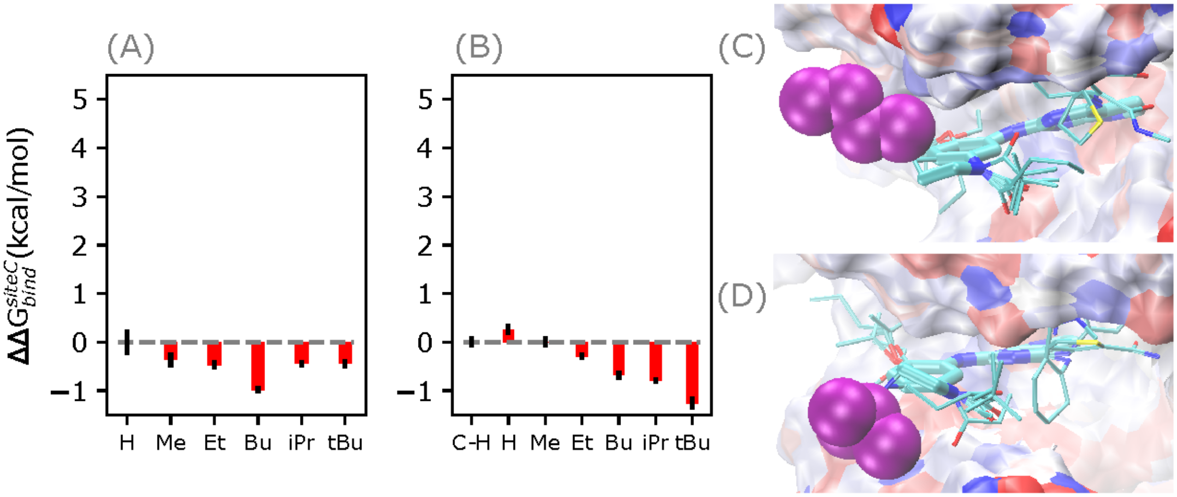

__Figure 5.__ (A-B) Relative binding free energies when substituent at site C is varied, while site B is carbon (A) or nitrogen (B). The results were obtained from simulation set II and III, respectively, and can be found in __Table S1__ as well. (C) Illustration of how butyl (C) at site C fits the binding pocket in TSSK1B protein when site B is carbon. (D) Illustration of how tert-butyl at site C fits the binding pocket in TSSK1B protein when site B is nitrogen. See __Figure 4__ caption for coloring scheme.

Similar to sites A and C, site D is also largely solvent exposed (see **Figure 6C-D** for example). Moreover, there are three negatively charged residues (Asp 97, Glu 100, and Asp 140) near site D, which makes hydrophobic substituents, including methyl, ethyl, butyl, isopropyl, and tert-butyl, generally unfavorable. Interestingly, the covalent geometry of the common core is slightly different when site B is changed from carbon to nitrogen, which then affects the orientation of site D substituent. When site B is carbon, even a small hydrophobic group, like methyl, can be close to the negatively charged residues, thus leading to reduced binding affinity (**Figure 6A**). When site C is nitrogen, larger/bulkier hydrophobic groups are more unfavorable (**Figure 6B**). For this reason, we suggest it's probably better to keep site B and D the same as that in the original KQQ molecule (**Figure 2**). The predicted pKa of the site D group in the KQQ molecule is 7.7 based on MolGpKa [45], and the pKa of trimethylamine is 9.8 [46]. Therefore, molecule KQQ may be protonated at site D with certain probabilities and form favorable interactions with these three negatively charged protein residues.

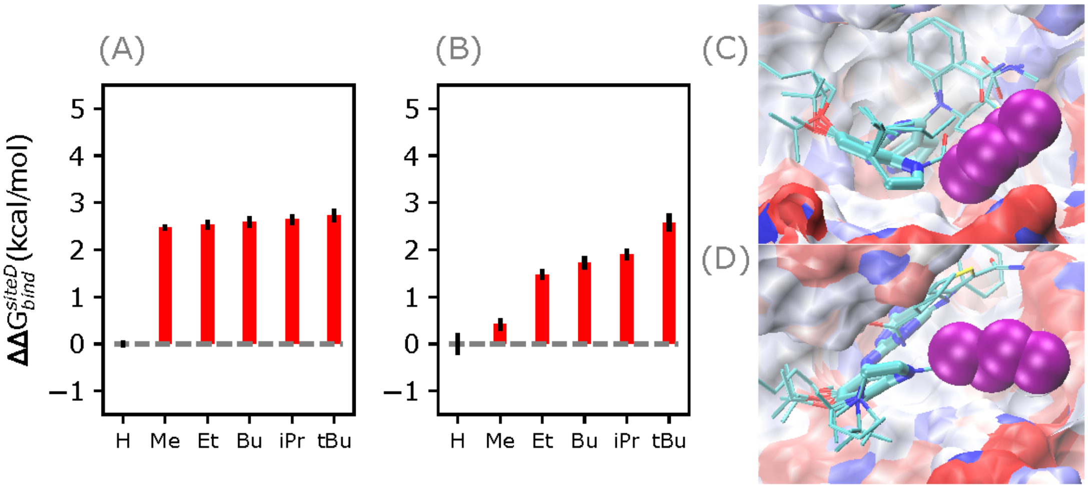

__Figure 6.__ (A-B) Relative binding free energies when substituent at site D is varied, while site B is either carbon (A) or nitrogen (B). The results were obtained from simulation set II and III,

respectively, and can be found in ***Table S1*** *as well. (C-D) Illustration of how butyl at site D fits the binding pocket in TSSK1B protein when site B is either carbon (C) or nitrogen (D). See* ***Figure 4*** *caption for coloring scheme.*

Adding more polar groups at site E further increases binding affinity

Although the three substituents at site E only showed small differences in binding free energies (see **Table 2**), site E is less solvent exposed and has multiple interactions with the protein. As illustrated in **Figure 7B**, the aromatic ring in these three substituents could form stable hydrophobic contacts with the protein. Also, depending on the orientation of the substituent, additional polar interactions can be observed (**Figure 7C**). Taken together, we anticipate that further modifying site E may help design better inhibitors of TSSK1B.

Therefore, we changed the site E substituent by adding more hydrophobic or polar groups and examined the effects on binding affinity. Here, we performed another set of MSλD simulations (simulation set IV in **Table I**) by modifying position Y of substituent 1 at site E (see **Figure 7A**). As shown in **Figure 7A**, adding a polar group, like an amine or hydroxyl, further lowers the binding free energy. This polar group can form contacts with polar/charged residues in TSSK1B, including Asp 154, Asp 140, and Lys 41. Further experimental studies will be needed to validate our predictions.

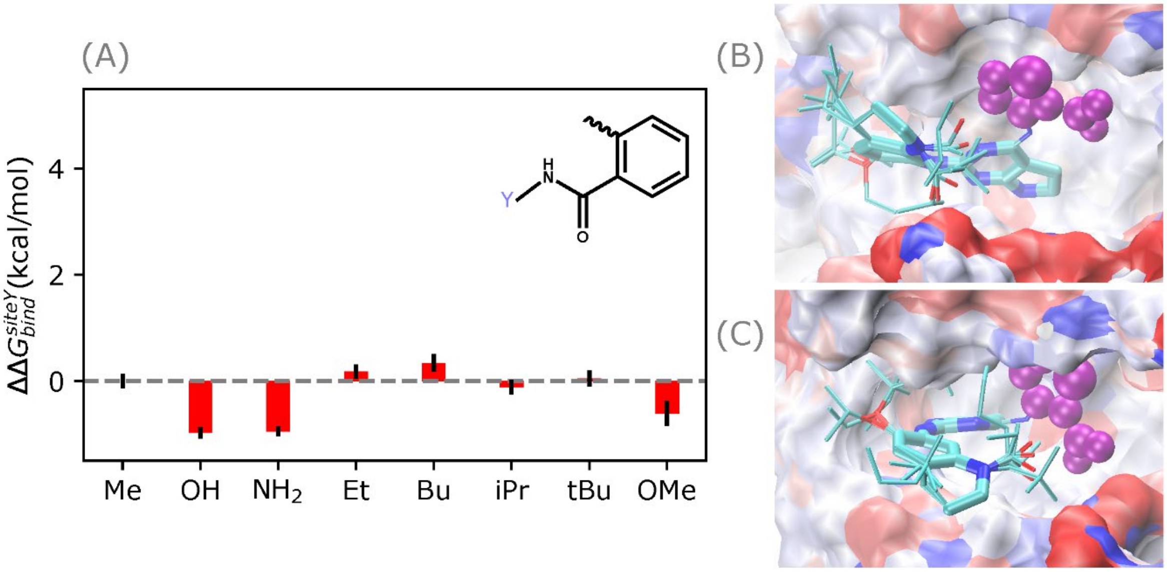

Figure 7. *(A) Relative binding free energies when substituent 1 at site E is further modified. The results were obtained from simulation set IV. (B-C) Illustration of how substituent 3 at site E fits the binding pocket of TSSK1B protein in simulation set II. See* ***Figure 4*** *caption for coloring scheme. Note that the radii of VDW spheres have been scaled to 50% to better show the protein surface.*

Conclusion

Being able to rapidly explore the large chemical space of a combinatory library and establish the QSAR relationship is very important in the lead optimization process of drug design. Unlike other popular free energy methods like FEP or TI, MSλD is a novel method that allows one to compute the relative free energy of many compounds within a single simulation and shows excellent scalability to the total number of compounds modeled. In the present study, we used MSλD to explore a large combinatorial chemical space by calculating the relative binding free energies of 1296 inhibitors to kinase TSSK1B, a validated target for male contraception. For this system, we found that the sites A, C, D and E in the small molecule scaffold are independent of each other, which allowed us to examine each site individually and compute relative binding free energies of these ligands using an additive model. This significantly reduced the required computational resources with no loss of statistical precision. Note that the lack of cooperativity between two sites in this study may not necessarily be generalized to other systems. It is possible that a ligand may take advantage of cooperative effects between different sites to achieve highest possible binding affinity. Depending on the target of interest and the ligand scaffold, such cooperative effective may need to be considered. Based on our simulations, we have also identified that sites A, C and D are largely solvent exposed and might be less sensitive to modifications. However, site E in the ligand is able to form both hydrophobic and polar interactions with the protein, and further modifications, such as adding more polar groups, may lead to increased binding affinity. Based on our calculations, we predict that by modifying sites A of the ligand KQQ to butyl, and position Y of substituent 1 at site E to an amine or hydroxyl group, we could further increase the binding affinity by ~1.5 kcal/mol.

Supplementary Information

Comparison between different homology models of TSSK1B, probability of forming contacts between molecule KQQ and the C-terminal tail of TSSK1B, reversible transitions in the alchemical space, experimental measurements of IC_{50}, optimization of dihedral potentials, verification of site independence for simulation set II, relative binding free energy values estimated using the additive model.

Acknowledgment

The authors gratefully acknowledge funding from GM130587. The authors also thank Ryan L. Hayes, Luis F. Cervantes and other members in Brooks group for stimulating discussions.

Author Contributions

C. L. B. and M. B. S., conception and design of the study; X. L., performing the simulations and data analysis; P. K. T., performing the experiments and data analysis; X. L., P. K. T., M. B. S. and C. L. B., interpretation of data, drafting and revising the manuscript.

References

[1] Zwanzig RW. High-Temperature Equation of State by a Perturbation Method .1. Nonpolar Gases. J Chem Phys. 1954;22:1420-6.
[2] Wang L, Berne BJ, Friesner RA. On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. Proc Natl Acad Sci U S A. 2012;109:1937-42.
[3] Straatsma TP, Berendsen HJC. Free-Energy of Ionic Hydration - Analysis of a Thermodynamic Integration Technique to Evaluate Free-Energy Differences by Molecular-Dynamics Simulations. J Chem Phys. 1988;89:5876-86.
[4] Wang L, Wu YJ, Deng YQ, 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, Murcko M, Frye L, Farid R, Lin T, Mobley DL, Jorgensen WL, Berne BJ, Friesner RA, Abel R. 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;137:2695-703.
[5] Steinbrecher TB, Dahlgren M, Cappel D, Lin T, Wang LL, Krilov G, Abel R, Friesner R, Sherman W. Accurate Binding Free Energy Predictions in Fragment Optimization. J Chem Inf Model. 2015;55:2411-20.
[6] Rufa DA, Macdonald HEB, Fass J, Wieder M, Grinaway PB, Roitberg AE, Isayev O, Chodera JD. Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning/molecular mechanics potentials. BioRxiv. 2020.
[7] Knight JL, Leswing K, Bos PH, Wang L. Impacting Drug Discovery Projects with Large-Scale Enumerations, Machine Learning Strategies, and Free-Energy Predictions. Free Energy Methods in Drug Discovery: Current State and Future Directions: ACS Publications; 2021. p. 205-26.
[8] Schindler CEM, Baumann H, Blum A, Bose D, Buchstaller HP, Burgdorf L, Cappel D, Chekler E, Czodrowski P, Dorsch D, Eguida MKI, Follows B, Fuchss T, Gradler U, Gunera J, Johnson T, Lebrun

CJ, Karra S, Klein M, Knehans T, Koetzner L, Krier M, Leiendecker M, Leuthner B, Li LW, Mochalkin I, Musil D, Neagu C, Rippmann F, Schiemann K, Schulz R, Steinbrecher T, Tanzer EM, Lopez AU, Follis AV, Wegener A, Kuhn D. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. J Chem Inf Model. 2020;60:5457-74.

[9] Raman EP, Paul TJ, Hayes RL, Brooks CL, III. Automated, Accurate, and Scalable Relative Protein-Ligand Binding Free-Energy Calculations Using Lambda Dynamics. J Chem Theory Comput. 2020;16:7895-914.

[10] Shirts MR, Mobley DL, Brown SP. Free-energy calculations in structure-based drug design. Drug design: structure-and ligand-based approaches. 2010:61-86.

[11] Pohorille A, Jarzynski C, Chipot C. Good practices in free-energy calculations. J Phys Chem B. 2010;114:10235-53.

[12] Liu S, Wu Y, Lin T, Abel R, Redmann JP, Summa CM, Jaber VR, Lim NM, Mobley DL. Lead optimization mapper: automating free energy calculations for lead optimization. J Comput Aided Mol Des. 2013;27:755-70.

[13] Kong XJ, Brooks CL, III. lambda-Dynamics: A new approach to free energy calculations. J Chem Phys. 1996;105:2414-23.

[14] Knight JL, Brooks CL, III. Multisite lambda Dynamics for Simulated Structure-Activity Relationship Studies. J Chem Theory Comput. 2011;7:2728-39.

[15] Vilseck JZ, Sohail N, Hayes RL, Brooks CL, III. Overcoming Challenging Substituent Perturbations with Multisite lambda-Dynamics: A Case Study Targeting beta-Secretase 1. J Phys Chem Lett. 2019;10:4875-80.

[16] Vilseck JZ, Armacost KA, Hayes RL, Goh GB, Brooks CL, III. Predicting Binding Free Energies in a Large Combinatorial Chemical Space Using Multisite lambda Dynamics. J Phys Chem Lett. 2018;9:3328-32.

[17] The Human Protein Atlas. Available from: https://www.proteinatlas.org/.

[18] Kadiyska T, Tourtourikov I, Dabchev K, Madzharova D, Tincheva S, Spandidos DA, Zoumpourlis V. Role of testis-specific serine kinase 1B in undiagnosed male infertility. Mol Med Rep. 2022;25.

[19] National Institutes of Health, Illuminating the Druggable Genome, Understudied Proteins. Available from: https://commonfund.nih.gov/IDG/understudiedproteins.

[20] H Drewry D, M Willson T, J Zuercher W. Seeding collaborations to advance kinase science with the GSK Published Kinase Inhibitor Set (PKIS). Curr Top Med Chem. 2014;14:340-2.

[21] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46:W296-W303.

[22] Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9.

[23] Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5:725-38.

[24] Yang JY, Yan RX, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2015;12:7-8.

[25] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millan C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker

D. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373:871-6.

[26] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583-9.

[27] Vilseck JZ, Cervantes LF, Hayes RL, Brooks CL, III. Optimizing Multisite lambda-Dynamics Throughput with Charge Renormalization. J Chem Inf Model. 2022;62:1479-88.

[28] Homan KT, Larimore KM, Elkins JM, Szklarz M, Knapp S, Tesmer JJG. Identification and Structure-Function Analysis of Subfamily Selective G Protein-Coupled Receptor Kinase Inhibitors. ACS Chem Biol. 2015;10:310-9.

[29] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, MacKerell AD. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. J Comput Chem. 2010;31:671-90.

[30] Yu WB, He XB, Vanommeslaeghe K, MacKerell AD. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. J Comput Chem. 2012;33:2451-68.

[31] Olsson MH, Sondergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput. 2011;7:525-37.

[32] Sondergaard CR, Olsson MH, Rostkowski M, Jensen JH. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. J Chem Theory Comput. 2011;7:2284-95.

[33] Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmuller H, MacKerell AD. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods. 2017;14:71-3.

[34] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER. J Chem Phys. 1983;79:926-35.

[35] Hayes RL, Armacost KA, Vilseck JZ, Brooks CL, III. Adaptive Landscape Flattening Accelerates Sampling of Alchemical Space in Multisite lambda Dynamics. J Phys Chem B. 2017;121:3626-35.

[36] Hayes RL, Vilseck JZ, Brooks CL, III. Approaching protein design with multisite lambda dynamics: Accurate and scalable mutational folding free energies in T4 lysozyme. Protein Sci. 2018;27:1910-22.

[37] Knight JL, Brooks CL, III. Applying Efficient Implicit Nongeometric Constraints in Alchemical Free Energy Simulations. J Comput Chem. 2011;32:3423-32.

[38] Brooks BR, Brooks CL, III, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The Biomolecular Simulation Program. J Comput Chem. 2009;30:1545-614.

[39] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comput Chem. 1983;4:187-217.

[40] Hynninen AP, Crowley MF. New Faster CHARMM Molecular Dynamics Engine. J Comput Chem. 2014;35:406-13.

[41] Darden T, York D, Pedersen L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. J Chem Phys. 1993;98:10089-92.

[42] Vangunsteren WF, Berendsen HJC. Algorithms for Macromolecular Dynamics and Constraint Dynamics. Mol Phys. 1977;34:1311-27.

[43] Gapsys V, Perez-Benito L, Aldeghi M, Seeliger D, Van Vlijmen H, Tresadern G, de Groot BL. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. Chem Sci. 2020;11:1140-52.

[44] Hayes RL, Vilseck JZ, Brooks CL, III. Addressing Intersite Coupling Unlocks Large Combinatorial Chemical Spaces for Alchemical Free Energy Methods. J Chem Theory Comput. 2022;18:2114-23.

[45] Pan X, Wang H, Li C, Zhang JZ, Ji C. MolGpka: A web server for small molecule p K a prediction using a graph-convolutional neural network. J Chem Inf Model. 2021;61:3159-65.

[46] Settimo L, Bellman K, Knegtel RMA. Comparison of the Accuracy of Experimental and Predicted pKa Values of Basic and Acidic Compounds. Pharm Res. 2014;31:1082-95.