Large-scale assessment of binding free energy calculations in active drug discovery projects

Christina EM Schindler,[*,†] Hannah Baumann,[†,‡] Andreas Blum,[†] Dietrich Böse,[†] Hans-Peter Buchstaller,[†] Lars Burgdorf,[†] Daniel Cappel,[¶] Eugene Chekler,[§] Paul Czodrowski,[†,‖] Dieter Dorsch,[†] Merveille K.I. Eguida,[†,⊥] Bruce Follows,[§] Thomas Fuchß,[†] Ulrich Grädler,[†] Jakub Gunera,[†] Theresa Johnson,[§] Catherine Jorand Lebrun,[§] Srinivasa Karra,[§] Markus Klein,[†] Lisa Kötzner,[†] Tim Knehans,[†,#] Mireille Krier,[†] Matthias Leiendecker,[†] Birgitta Leuthner,[†] Liwei Li,[§] Igor Mochalkin,[§,△] Djordje Musil,[†] Constantin Neagu,[§] Friedrich Rippmann,[†] Kai Schiemann,[†] Robert Schulz,[†,@] Thomas Steinbrecher,[¶] Eva-Maria Tanzer,[§] Andrea Unzue Lopez,[†] Ariele Viacava Follis,[§] Ansgar Wegener,[†] and Daniel Kuhn[*,†]

E-mail: christina.schindler@merckgroup.com; daniel.kuhn@merckgroup.com

[*]To whom correspondence should be addressed

[†]Merck KGaA, Frankfurter Str. 250, 64293 Darmstadt, Germany

[‡]Present address: Department of Pharmaceutical Sciences, University of California, Irvine, California 92697, USA

[¶]Schrödinger GmbH, Q7 23, 68161 Mannheim, Germany

[§]EMD Serono Research and Development Institute Inc., 45A Middlesex Turnpike, Billerica, MA 01821, USA

[‖]Present address: Faculty of Chemistry and Chemical Biology, TU Dortmund University, Otto-Hahn-Strasse 6, 44227 Dortmund, Germany

[⊥]Present address: Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Universite de Strasbourg, 74 route du Rhin, 67400 Illkirch-Graffenstaden, France

[#]Present address: Schrödinger GmbH, Q7 23, 68161 Mannheim, Germany

[@]Institue of Pharmacy, Freie Universität Berlin, Königin-Luise-Straße 2+4, 14195 Berlin, Germany

[△]Present address: Vertex Pharmaceuticals, 3215 Merryfield Row, San Diego, CA 92121, USA

Abstract

Accurate ranking of compounds with regards to their binding affinity to a protein using computational methods has long been of great interest to pharmaceutical research. Physics-based free energy calculations are regarded as the most rigorous way to estimate binding affinity. In recent years, many retrospective studies carried out both in academia and industry have demonstrated its potential. Here, we present the results of large-scale *prospective* application of the FEP+ method in active drug discovery projects in an industrial setting at Merck KGaA, Darmstadt, Germany. We compare these prospective data to results obtained on a new diverse benchmark of pharmaceutically relevant targets. Our results offer insights into the challenges faced when using free energy calculations in real-life drug discovery projects and identify limitations that could be tackled by future method development.

Introduction

Identifying ligands that bind with high affinity to a target protein, while balancing other ligand properties relevant to safety and biological efficacy, is the goal of small-molecule drug discovery projects. To support this challenging enterprise, accurate prediction of protein-ligand binding free energies has long been a goal of computer-aided drug design (CADD). Molecular dynamics based free energy calculations are considered the most rigorous approach to this problem. Yet, until recently, wide-spread usage was hindered by high computational costs, limitations of sampling algorithms and limitations in (small molecule) force fields. In addition, from an industry perspective, using free energy calculations was considered challenging due to the limited automation, throughput and robustness of the protocols.

Over the last 10 years, however, there has been tremendous progress in sampling,[1–5] force field development,[6–15] throughput[16–24] and automation.[5,11,25–28] As a result, many pharmaceutical companies have adopted relative free energy calculations and in particular the Schrödinger FEP+ workflow[5] as a new computational tool to support their drug discov-

ery efforts. In 2016, we started a large initiative at Merck KGaA, Darmstadt, Germany to thoroughly assess the prediction accuracy and identify the best use cases of free energy calculations. In this initiative, we aimed to apply FEP+ in all suitable in-house active drug discovery projects prospectively. There were three main reasons for this particular setup of testing a new method in a real-life prospective application. First, since a large initial assessment of the FEP+ method in 2015,[5] many features had been added to the method, e.g., calculating free energies for transformations involving ring openings and net charge changes.[29,30] These challenging types of transformations are often used in drug discovery projects but were not present in the previous benchmark. Second, blind assessment of computational tools – as in the form of community wide prediction challenges such as CASP,[31] CAPRI,[32] SAMPL[33] and D3R Grand Challenges[34] – has been demonstrated to provide a more realistic picture of the accuracy and limitations of a given method. Therefore, in this initiative we focused on applying FEP+ prospectively; i.e., in a blind manner. Third, it was unclear how much time constraints, limitations on resources and information gaps that are prevalent in real-life drug discovery projects affect the performance of the method relative to what was reported previously in the literature.[5,14,29,30]

Here, we present retrospective and prospective data collected in 17 in-house drug discovery projects over three years. We describe the general workflow we established for using free energy calculations in projects and report the performance of the FEP+ method on in-house targets. We further present data collected from a new diverse benchmark and discuss key learnings for domain of applicability, project impact and limitations of the method.

Results and discussion

Free energy calculations have emerged as an accurate binding affinity prediction method that could potentially accelerate hit-to-candidate optimization. However, it is not clear how accurate this method is when applied under the constraints of an industry setting and how

it can best contribute to compound design and the related multi-parameter optimization process. Therefore, in 2016, a large initiative was launched at Merck KGaA, Darmstadt, Germany to thoroughly evaluate and prospectively benchmark the FEP+ technology in active projects. From 2016 to 2019, we applied FEP+ prospectively to 12 targets and 23 chemical series performing over 35,000 individual perturbation calculations. We finally obtained valid predictions for over 6,000 chemical entities. More than 400 blindly predicted and novel molecules were synthesized and tested. This yielded a large set of prospective data providing a realistic assessment of the method's accuracy in a typical small molecule drug discovery working environment. In addition to benchmarking, this initiative also enabled us to define best practices and explore optimal use cases.

Free energy calculation workflow in projects

Throughout the last three years, we established a workflow for deploying free energy calculations in projects. The process is shown schematically in Figure 1. First, we assess the general feasibility of using FEP for a given target and chemical series of interest by collecting available structural data and experimentally determined binding affinities. At this stage, we typically require at least one well-resolved co-crystal structure with one representative of the series of interest. This strict requirement is based on our experiences in three projects where we tried to use homology models in the absence of an X-ray structure. In contrast to previous successful applications of FEP with homology models,[35–37] we obtained unsatisfactory prediction accuracy retrospectively in all three cases. In two of the three projects, we later had access to a co-crystal structure and were then able to successfully complete the validation phase (see below). Once we have sufficient structural data available, we collect a data set of congeneric ligands with experimental binding affinities (at least 10 ligands, preferably 20) and all available information on the biochemical and biophysical assays (e.g., protein construct, buffer conditions, presence of co-factors etc.). If the ligand data set is large, it can be split according to the R-groups of the molecule that are modified to identify

potentially different accuracy for predictions in different parts of the binding site. Then based on these data sets, retrospective free energy calculations are performed in order to compare predicted to experimentally determined binding affinities. We typically refer to these retrospective computational experiments as validation studies. The accuracy of the free energy calculations is assessed by calculating the root-mean-square error (RMSE) between relative predicted affinities $\Delta\Delta G_{\mathrm{pred}}$ and relative experimental affinities $\Delta\Delta G_{\mathrm{exp}}$ for all ligand pairs in the set. If available, predicted affinities are compared to experimental results from different assays. Typically, during the validation phase, different input structures and settings such as sampling time are evaluated in order to find the best setup for later prospective calculations. In practice, time constraints often limit this phase to evaluating typically 3 possible model systems. In case of large outliers ($|\Delta G_{\mathrm{pred}} - \Delta G_{\mathrm{exp}}| > 2$ kcal/mol), detailed analysis is done in order to understand the underlying causes. We usually consider a validation study as successful if we achieve an RMSE smaller than 1.3 kcal/mol and if we are able to sufficiently explain large outliers if present. If the dynamic range of the data set is suitable, FEP predictions should also yield good ranking. In reality, however, we frequently encountered the situation that only a data set with a limited dynamic range was available at the time of the validation study, which made model evaluation challenging.[38] After successfully completing validation stage, the FEP project moves into production mode and prospective calculations are performed for compound ideas. These ideas have to be sufficiently similar to the validated chemical series. In case of a new chemical series and new available structural information, a new validation study has to be performed. We closely monitor the prospective accuracy of the predictions throughout the project and track which predicted compounds have been synthesized. All predicted affinities and structures are stored in a database. Using an automated workflow, we periodically check our in-house database to detect whether predicted compounds have been synthesized and tested in the meantime. Structure matching is done either based on exact structure matches, matching tautomers or matching without considering stereochemistry to maximize the number of data

collected. In our experience, this constant monitoring is necessary to initially build trust in the project team and later detect when the chemical matter has moved outside of the domain of applicability of the model.

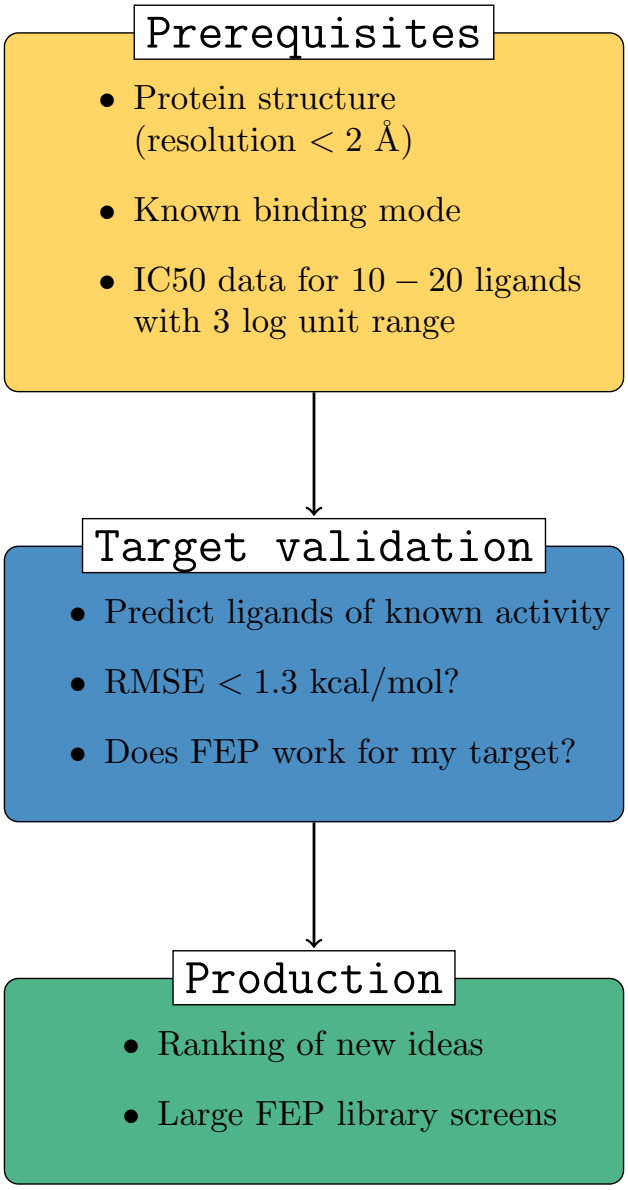

Figure 1 Deploying free energy calculations in drug discovery projects at Merck KGaA, Darmstadt, Germany.

FEP feasibility and validation results in in-house drug discovery projects

Over the course of three years, we evaluated 28 targets for general FEP feasibility (Figure 2 A). We performed validation studies for 17 targets and 44 chemical series and progressed 12 targets and 23 chemical series to prospective calculations (as mentioned before, initially our criteria for considering a validation as successful were more lenient). The major reason for not being able to perform a validation study for a given target was the lack of relevant structural data (10 targets). For two targets, all validation studies were considered unsuccessful and for another two targets, the projects were stopped shortly after an initial FEP feasibility evaluation was done (portfolio category). Some chemical series were not progressed to production mode despite good validation results, as they had been deprioritized in the meantime by the project team for other reasons (however, other series from the same project entered production mode in contrast to the two FEP projects in the portfolio category, see above). Overall, we observed a relatively low attrition rate for potential FEP targets, once enough structural and binding affinity data became available to perform a validation study.

The RMSEs achieved in the validation study for 17 targets are shown in Figure 2 (B). In total, we were able to obtain high accuracy (RMSE $<$ 1 kcal/mol) and acceptable accuracy (RMSE $<$ 1.3 kcal/mol) predictions for 13 targets and 20 chemical series. At earlier stages in the initiative, we also accepted validation studies with an RMSE larger than 1.3 kcal/mol as successful and therefore progressed these series to production mode. During the course of our evaluation, we raised the requirements for a successful validation study, since lower accuracy (RMSE between 1.3 and 1.6 kcal/mol) in validation studies consistently led to even lower accuracy in a prospective setting. Prediction accuracy varied not only between different targets but also between different chemical series for the same target protein. Most notably, for one target, we obtained results for multiple chemical series covering the whole range from high accuracy (RMSE $<$ 1 kcal/mol) to low accuracy predictions (RMSE $>$ 2 kcal/mol) (Figure 2 (B), light pink color). Furthermore, when using FEP in active projects, we regularly

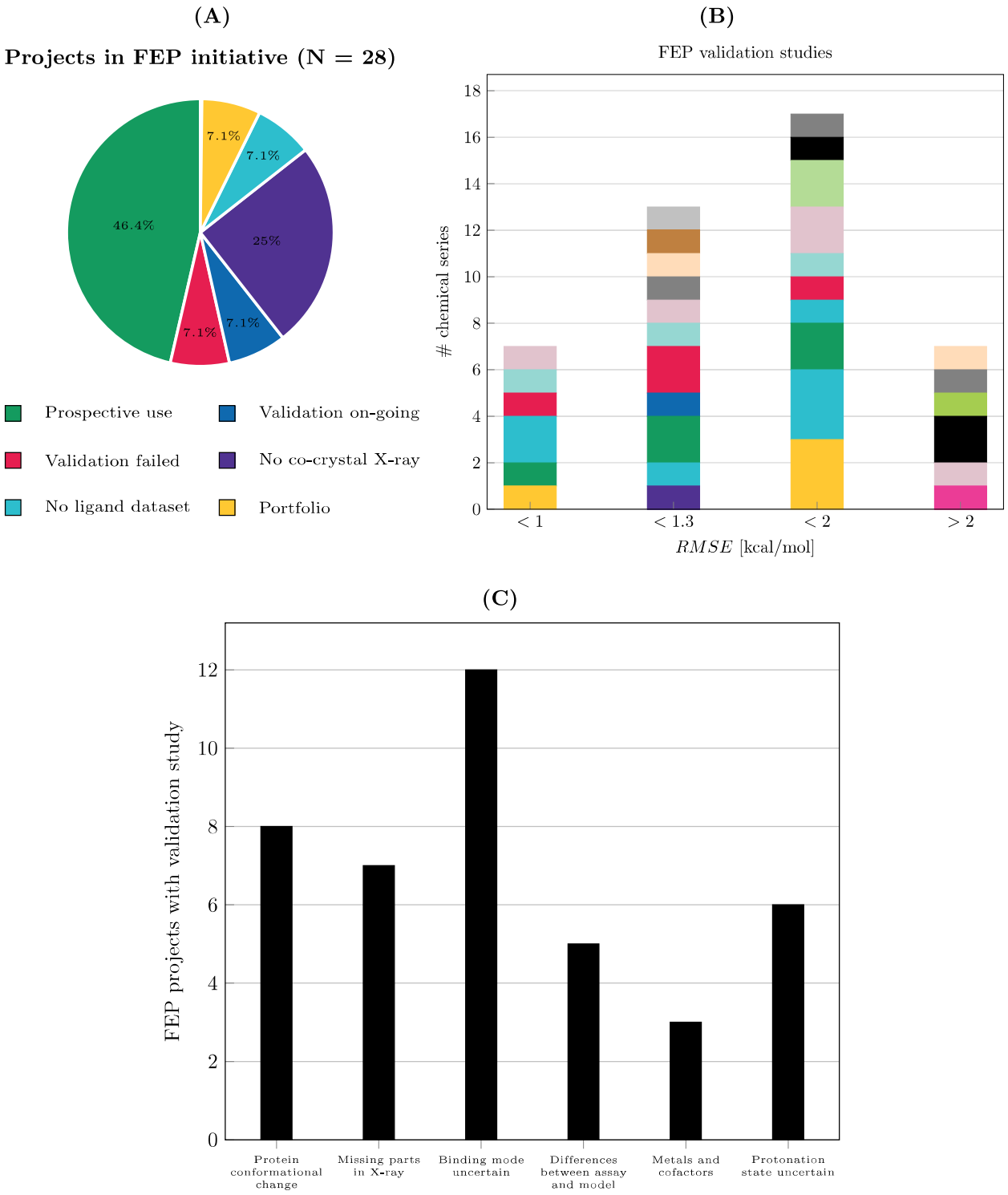

Figure 2 FEP feasibility, validation results and challenges in in-house projects. (A) Outcome of FEP evaluation on 25 targets. The main reason for not being able to use FEP in projects was the lack of structural data. (B) Results for validation studies using FEP+. Different colors represent different target receptors. For many targets, several chemical series were evaluated. The RMSE varies greatly depending not only on the target but also on the chemical series of interest. (C) Possible challenges for prediction accuracy are encountered in all projects.

faced challenges that might affect the accuracy of the method. A qualitative assessment of these challenges is shown in Figure 2 (C). Almost all projects had at least one aspect that might be problematic for applying free energy calculations (Figure 2 (C)): unsurprisingly, real-life drug discovery projects are not ideal case scenarios. The most common challenges we encountered in the projects where we attempted a validation study were uncertainties in the binding mode of at least a subset of the ligands and uncertainties in the protein structure due to suspected protein conformational change (71% and 47% respectively). In five projects, we found that the source of experimental data had an influence on the success of the validation study. In one case, we initially compared the predicted affinities to the output of a functional assay and found large deviations. However, when comparing the same predicted affinities to biophysical SPR data, we found good agreement and therefore decided to progress the series to production mode (Figure S1). On the one hand, it is expected that FEP results should in general correspond best to biophysical binding data and could display larger errors when compared to other types of assays. On the other hand, to have strong impact in projects, FEP predicted affinities should correlate well to the main assay that is used to drive compound optimization. Also, in many projects, there is only a limited amount of biophysical data available compared to data from biochemical or functional assays.

Note that here we listed *potential* aspects that may affect the accuracy of the predictions. However, we were not generally able to pinpoint the probability of success for a validation study to a specific feature. For example, we analyzed whether the resolution of the X-ray structure had an influence on the RMSE achieved in the validation study but found no correlation (data not shown).

Prospective FEP+ results for in-house projects

For 12 targets and 19 chemical series, it was possible to obtain a prospective data set consisting of at least five data points. The results for each target and series are shown in Figure 3 (see also Table S1). Compared to the results from the validation study, we observed a lower

accuracy in prospective applications as indicated by higher RMSE values. Note that here we measured accuracy comparing predicted free energy ΔG_{pred} to experimental affinity ΔG_{exp} instead of calculating RMSE on relative free energies $\Delta\Delta G$ as in the validation study, since the ligands did not originate from a single FEP map. For three (albeit small) data sets, we obtained high accuracy predictions (RMSE $<$ 1 kcal/mol), for four sets, medium accuracy (RMSE $<$ 1.5 kcal/mol; this is the accuracy we expect to see for a series that showed RMSE $<$ 1.3 kcal/mol in the validation study) and for another eight sets, acceptable accuracy predictions (RMSE $<$ 2 kcal/mol). Binding affinity prediction methods with moderate accuracy of RMSE $<$ 2 kcal/mol are considered useful for scoring larger libraries.[39] We were able to achieve this moderate accuracy level in 17 out of 19 chemical series from our prospective applications. For 13 out of the 19 sets, we either found good correlation (Kendall $\tau > 0.5$) or low error (RMSE $<$ 1.5 kcal/mol). In Figure 3, the targets are ordered chronologically. This data therefore also displays our "FEP learning curve". Indeed, the RMSE for the first five series is markedly higher than the RMSE calculated for the most recent five data sets (RMSE= $1.66^{1.92}_{1.38}$ kcal/mol for 175 ligands and RMSE = $1.35^{1.72}_{0.97}$ kcal/mol for 118 ligands respectively).

We investigated the reasons for the largest outliers in the data set. For the analysis, we reset the predicted affinities of all compounds that were predicted to be above the top or below the bottom of the assay to these values and then evaluated the resulting deviation from the experimental value. Based on this analysis, we still identified 23 compounds where the predicted affinity deviated by more than 3 kcal/mol from the experimental affinity. For four of these 23 outliers that originated from target 1/series 1, the validation study had previously shown larger errors for modifications in this part of the molecule. Based on the stricter criteria established later in the FEP initiative, we would not have progressed this project to prospective phase. Another four compounds from target 1/series 2 and target 1/series 3 displayed changes in aromatic heterocycles relative to the reference compound. These heterocycles might have not been well represented by the force field (the affinity of

these compounds was underestimated, see Figure S2 for examples). For three compounds from target 2/series 1 and target 5/series 1, we suspect that the changes relative to the molecule that was co-crystallized could have invoked protein conformational changes that were not captured in the simulations (the changes were made in a part of the binding site that displayed flexibility in the available crystal structures). For a group of five outliers from target 5/series 1, we suspect that an additional domain that was not part of the crystallization construct might have caused the sudden drop in potency when making modifications to the solvent-exposed side of the molecule (the five molecules were found to be inactive in the assay despite having relatively small modifications compared to highly active molecules). All of these molecules were over-predicted with FEP by more than 5 kcal/mol, contributing significantly to the overall RMSE of $2.16^{2.64}_{1.63}$ kcal/mol observed for this data set. When excluding these five molecules, the RMSE calculated over the remaining 83 compounds was reduced to $1.38^{1.66}_{1.12}$ kcal/mol. Finally, two of the outliers that originated from target 8 and target 9 had modifications to a sulfone or a sulfonamide. When we initially performed the FEP+ calculations for these moecules, we noticed that many modifications were predicted to be favorable when replacing the sulfone or the sulfonamide group (which already seemed "suspicious" at the time). Indeed, when we synthesized the best predicted compounds in these two projects, we found that the affinity was over-predicted by more than 4 kcal/mol. These examples might hint at a general problem of small molecule force fields in describing the properties of sulfones in relatively hydrophobic binding pockets.

In general, there are several reasons that could account for the larger errors in the prospective data sets compared to the validation studies. First, throughout the course of a project, new ideas tend to be decreasingly similar to the chemical space that was used in the validation study and decreasingly similar to the representative of the chemical series that was captured in the crystal structure. This naturally results in higher uncertainties in e.g., binding mode and protonation state of the ligands. In many cases, we tried to simulate either multiple binding modes, tautomers or charge states in the map when ranking the full set of

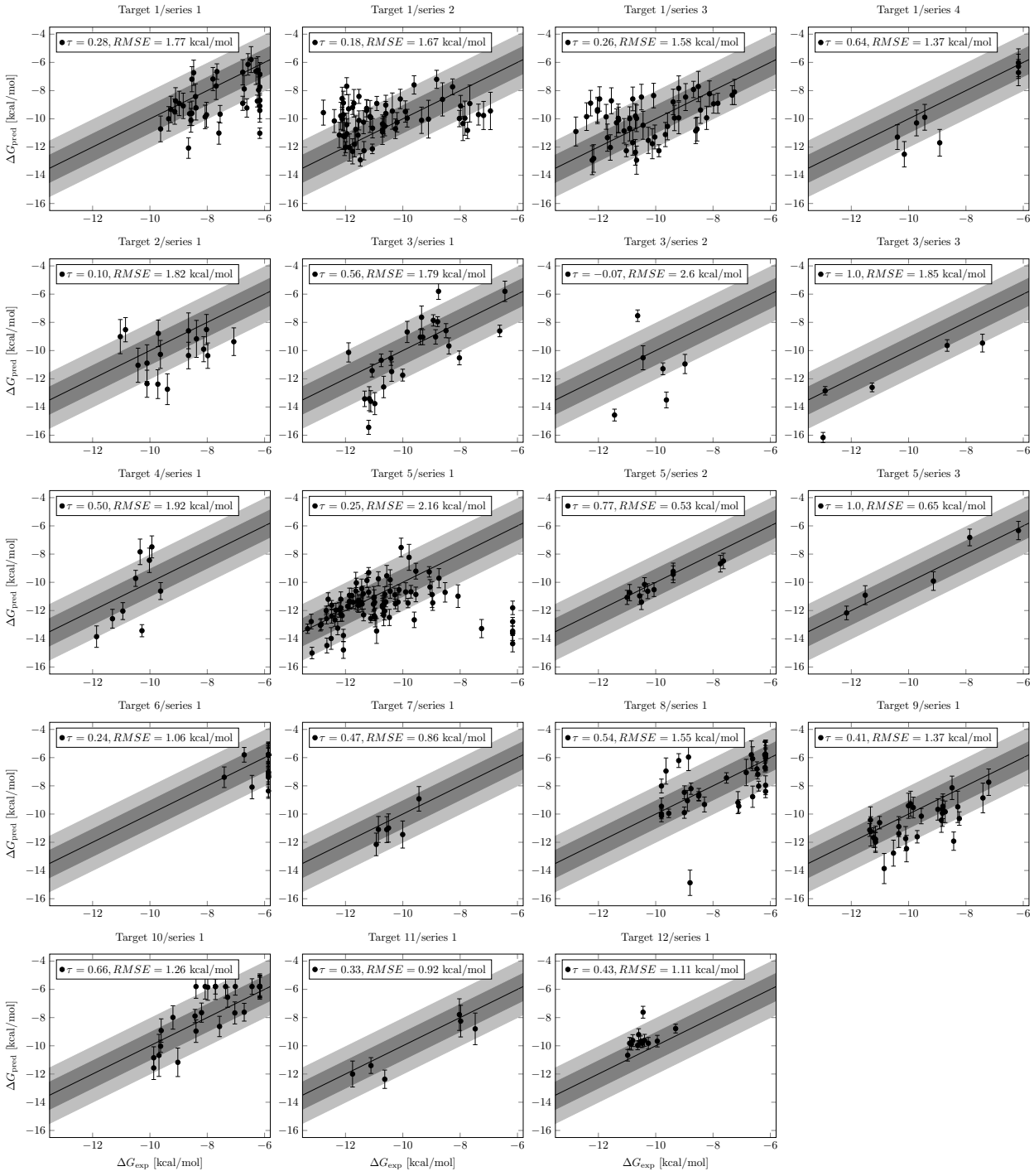

Figure 3 Prospective FEP+ results from 12 targets and 19 chemical series. Experimental affinities were converted to ΔG values using the equation $\Delta G_{\text{exp}} \approx k_B T \log \text{IC}_{50}$. FEP+ accuracy varies between different targets and chemical series. Predictions within 1 kcal/mol and 2 kcal/mol of the experimental affinity are highlighted by dark and light gray area respectively.

compounds. Alternatively, we first tried to predict the relevant state with FEP+ (binding pose FEP) and then ranked the whole set based on the predicted optimal state for each compound. Recently, there has been a possibility to include multiple protonation states in FEP+.[40] However, cases in which the protein changes its protonation state upon binding or cases where the ligand has a different pKa in the binding site than in the protein[41–43] still cannot be treated properly. Recently, there has been progress in the field of constant pH simulations,[44–48] yet at the moment, these methods appear too computationally demanding for an industry context. Second, we found that the diverse chemical matter in our in-house compound collection remained a challenge for small molecule force fields. For almost every new chemical series, we had to reparametrize some of the torsion potentials, even when using the new OPLS3e force field that includes a very high number of torsion potentials.[14] New concepts like the SmirnoFF force field developed by the Open Force Field Consortium[49] appear promising in helping to overcome the limitations of atom typing in current small molecule force fields. We anticipate that small molecule force fields will continue to improve and better capture the properties of different chemical groups in the future. This in turn will hopefully reduce errors in free energy calculations. Third, when using free energy calculations prospectively, we tend to focus on extreme predictions (e.g., the top-ranked compounds) that might be inherently more error-prone. This bias in focussing on extreme predictions can be softened by e.g. applying a selection bias correction.[50] However, in our hands this did not affect the largest outliers much. Finally, many transformations that were of high interest to the project teams were innately challenging for the method (e.g., from aromatic ring systems to aliphatic chains, charge changes, addition of new groups via flexible linkers). Especially, the addition of a new functional group with a flexible linker is difficult in terms of sampling, but it frequently occurs in the context of early hit optimization and in fragment optimization.

When comparing ranking by FEP+ to ranking obtained from Glide docking and Prime MMGB-SA scoring (see Experimental for details), we found that overall FEP+ performed

better than these standard SBDD methods (Table S1). Cohen's d for Kendall τ was 0.85 for comparison to Glide and 0.49 for Prime, indicating a medium and small effect size respectively.[51] Note that for comparison with Prime, Target 6/series 1 was not considered for calculating effect size and also that for technical reasons it was not possible to rank all ligands with the two other methods (e.g., some ligands failed to dock into the protein due to clashes or gave positives scores in Prime MM-GBSA and were therefore excluded from the analysis).

In four cases (target 1/series 4, target 4/series 1, target 5/series 3 and target 12/series 1), Prime MM-GBSA gave ranking with similar or slightly higher correlations than that of FEP+ based ranking. However, no quantitative agreement in terms of predicting absolute or relative free energies was found. Still, this highlights the opportunity to use simpler scoring methods for certain cases and focus with a computationally expensive method like FEP+ on those cases where other methods fail to rank the ligands.

We also compared FEP+ performance on our prospective in-house data sets to ranking with simple descriptors like molecular weight and log P (correlation statistics are shown in Table S2). FEP+ also outperformed these "null models" as indicated by a large and medium effect size (Cohen's d for Kendall τ $d = 1.03$ and 0.64 for comparing FEP+ with ranking by molecular weight and log P).[51] Glide docking performed no better than ranking by molecular weight and only showed a very small effect size with respect to ranking by log P (Cohen's d for $R^2 = -0.13$ and 0.07 respectively). Prime MM-GBSA scoring showed very small and small effect size (Cohen's d for $R^2 = 0.13$ and 0.33 when compared to ranking by molecular weight and log P respectively).[52]

Benchmark results

Based on our extensive experience with free energy calculations, we decided to construct a new benchmark consisting of eight challenging, recently published data sets with pharmaceutically relevant targets.[53–62] The benchmark comprises 264 ligands in total. It illustrates

many of the challenges we faced when using FEP+ in projects and reflects the typical type and size of chemical transformations during hit-to-lead and lead optimization. In contrast to a previously published benchmark,[5] the transformation include changes in the net charge and the charge distribution of the molecules as well as ring openings and core hopping (examples are shown in Figure 4). The ligand sets display a slightly increased range of structural diversity compared to the previous benchmark (average maximal mean pairwise Tanimoto similarity $0.79^{0.87}_{0.68}$ vs. $0.84^{0.92}_{0.76}$ using RDKit Daylight fingerprint with default settings). The FEP+ results are summarized in Table 1 (detailed plots for each set can be found in Figure 5). Overall, we achieve good correlation for these data sets (average $R^2 = 0.43^{0.64}_{0.25}$, average Kendall $\tau = 0.49^{0.64}_{0.33}$). The root-mean-square error for the relative affinities $\Delta\Delta G$ calculated over all ligand pairs in the map $\mathrm{RMSE}_{\mathrm{pair}}$ ranges between 1.2 and 2.1 kcal/mol. Given the challenges that are included in this data set, this is a remarkable achievement. However, compared to the accuracies reported earlier on a large benchmark set[5] and for internal drug discovery projects at Schrödinger (RMSE = 1.1 kcal/mol, Lingle Wang, personal communication), we see considerably lower accuracy. This lower accuracy is in line with the prospective accuracy found in our in-house projects (average $\mathrm{RMSE}_{\mathrm{pair}} = 1.68^{1.76}_{1.60}$ kcal/mol and average RMSE = $1.64^{1.97}_{1.26}$ kcal/mol respectively). When analyzing the error for the different types of transformations (Figure S3), we found that transformations involving changes of the net charge or the charge location or changes to the core/scaffold of the molecule, showed lower accuracy. These transformations are already considered as inherently more difficult by the FEP+ software and special settings for sampling are used.[29,30] Still, these transformations display larger errors. Interestingly, for the remaining transformations, we could not observe a strong dependency of the error on the size of the transformation; i.e., the number of heavy atoms that are changed.

It is interesting and quite sobering to note that based on our guidelines for a successful validation study (good correlation and $\mathrm{RMSE}_{\mathrm{pw}} < 1.3$ kcal/mol), we would have only been able to progress one out of the eight chemical series in the benchmark to production mode

based on the data obtained for the default setting of 5 ns smapling time. Increasing the simulation time from 5 ns to 20 ns per λ window decreased the RMSE for seven out of eight targets by more than 0.5 kcal/mol and decreased the average RMSE_{pw} by 0.13 kcal/mol (average RMSE = $1.51^{1.74}_{1.44}$ kcal/mol, see Table S3). This increase in quantitative agreement between predicted and experimental affinities had however no effect on ranking/correlation (average Kendall $\tau = 0.49^{0.64}_{0.32}$). Nevertheless, based on the 20 ns data, we would have been able to progress three out of the eight targets to production mode.

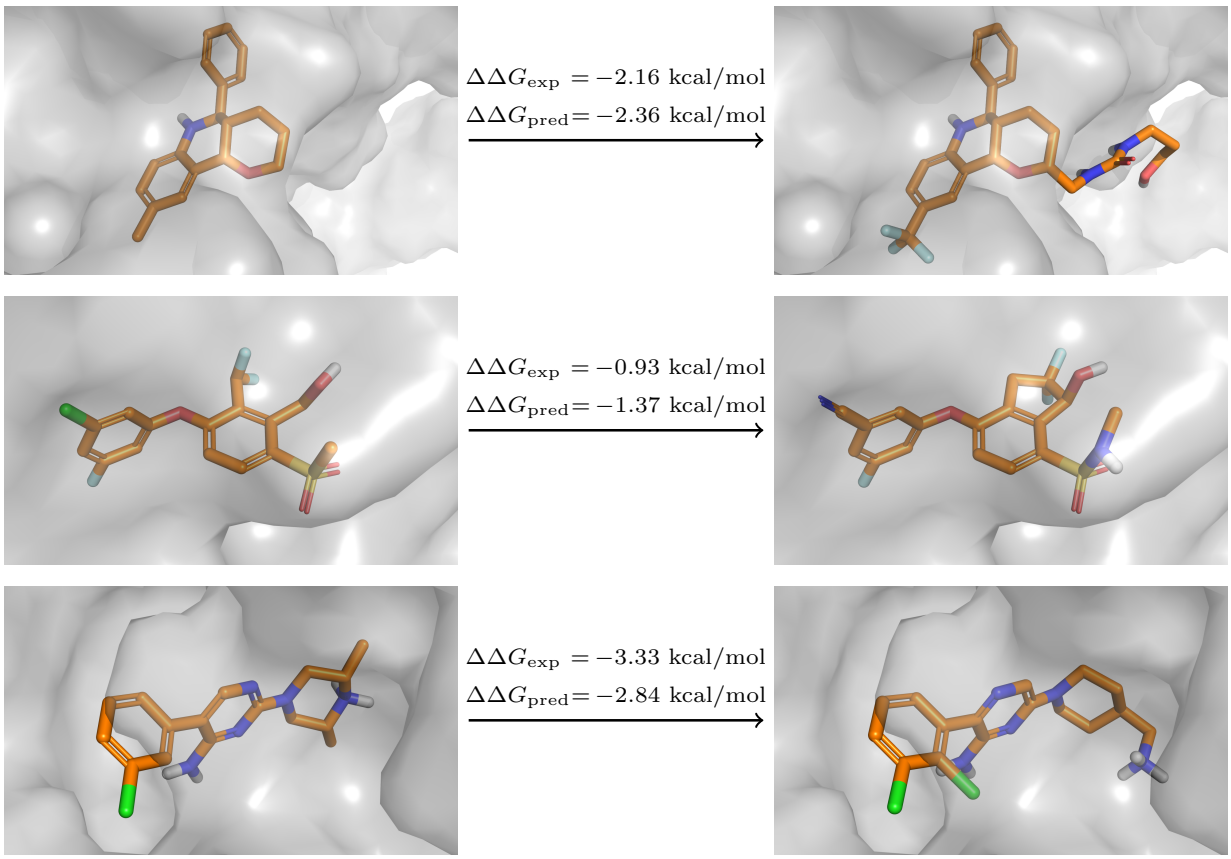

Figure 4 Examples of different types of transformations in the new benchmark set. (A) Addition of flexible chain in Eg5 leads to increased binding affinity. (B) Ring closure transformation in HIF-2α. FEP correctly predicts an increase in potency, likely due to the reduced ligand flexibility . (C) Shift of charged amine in SHP-2.

We discuss the c-Met FEP+ results as one example in more detail. The accuracy on this data set was moderate ($\text{RMSE}_{\text{pw}} = 1.43^{1.51}_{1.34}$ kcal/mol), but the predicted ΔG values show good correlation and ranking when compared to the experimental affinities (Figure 6

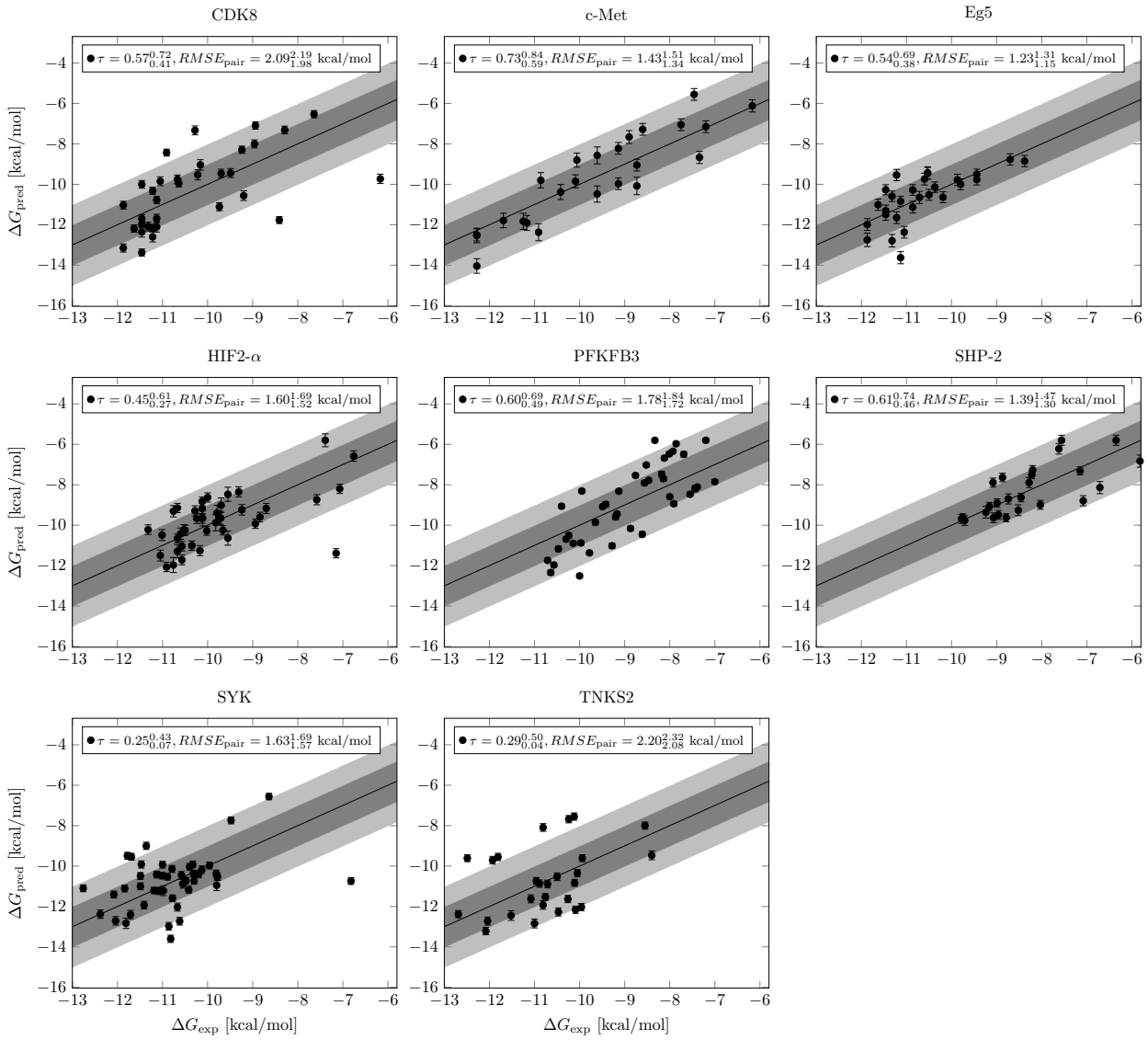

Figure 5 FEP+ results on literature curated benchmark. For each dataset, experimental and predicted ΔG values are shown. Calculations were run for 5 ns per window. Experimental affinities were converted to ΔG values using the equation $\Delta G_{\text{exp}} \approx k_B T \log \text{IC}_{50}$. Predictions within 1 kcal/mol and 2 kcal/mol of the experimental affinity are highlighted by dark and light gray area respectively.

Table 1 Results for ranking by FEP+, Glide and Prime MM-GBSA on literature curated benchmark set. Calculations were run for 5 ns per window (default sampling time). Pairwise RMSE was calculated for predicted and experimental $\Delta\Delta G$ values of all ligand pairs. Averages were calculated weighted by the number of the ligands in the respective data set. Pairwise RSME for Prime is shown for reference only (Prime scores typically range between -60 and -80). For each performance metric, confidence intervals were estimated via bootstrap sampling.

			FEP+				Glide				Prime			
Protein	N	PDB ID	R^2	ρ	τ	RMSE_{pw}	R^2	ρ	τ	RMSE_{pw}	R^2	ρ	τ	RMSE_{pw}
CDK8	33	5HNB	$0.38^{0.67}_{0.18}$	$0.74^{0.86}_{0.56}$	$0.57^{0.72}_{0.41}$	$2.09^{2.19}_{1.98}$	$0.0^{0.14}_{0.0}$	$0.13^{0.44}_{-0.19}$	$0.1^{0.33}_{-0.13}$	$2.49^{2.64}_{2.33}$	$0.6^{0.76}_{0.47}$	$0.82^{0.91}_{0.66}$	$0.64^{0.77}_{0.5}$	$7.03^{7.45}_{6.6}$
c-Met	24	4R1Y	$0.81^{0.89}_{0.7}$	$0.88^{0.94}_{0.74}$	$0.73^{0.84}_{0.59}$	$1.43^{1.51}_{1.34}$	$0.0^{0.18}_{0.0}$	$0.13^{0.48}_{-0.25}$	$0.1^{0.38}_{-0.2}$	$3.01^{3.19}_{2.82}$	$0.36^{0.63}_{0.14}$	$0.64^{0.83}_{0.34}$	$0.47^{0.67}_{0.24}$	$5.96^{6.38}_{5.53}$
Eg5	28	3L9H	$0.5^{0.69}_{0.33}$	$0.72^{0.84}_{0.52}$	$0.54^{0.68}_{0.39}$	$1.23^{1.31}_{1.15}$	$0.0^{0.15}_{0.0}$	$-0.08^{0.28}_{-0.4}$	$-0.03^{0.24}_{-0.28}$	$1.9^{2.0}_{1.79}$	$0.02^{0.12}_{0.0}$	$0.1^{0.38}_{-0.21}$	$0.06^{0.26}_{-0.14}$	$10.09^{10.65}_{9.54}$
HIF-2α	42	5TBM	$0.37^{0.65}_{0.1}$	$0.59^{0.77}_{0.36}$	$0.45^{0.61}_{0.27}$	$1.6^{1.69}_{1.52}$	$0.16^{0.36}_{0.04}$	$0.42^{0.61}_{0.18}$	$0.28^{0.44}_{0.12}$	$1.51^{1.57}_{1.45}$	$0.29^{0.51}_{0.09}$	$0.48^{0.65}_{0.27}$	$0.34^{0.48}_{0.2}$	$11.69^{12.17}_{11.21}$
PFKFB3	40	6HVI	$0.63^{0.75}_{0.5}$	$0.79^{0.86}_{0.67}$	$0.6^{0.69}_{0.49}$	$1.78^{1.84}_{1.72}$	$0.22^{0.46}_{0.08}$	$0.51^{0.71}_{0.25}$	$0.38^{0.56}_{0.2}$	$1.57^{1.65}_{1.49}$	$0.25^{0.44}_{0.1}$	$0.54^{0.7}_{0.33}$	$0.37^{0.51}_{0.22}$	$6.99^{7.29}_{6.68}$
SHP-2	26	5EHR	$0.5^{0.69}_{0.32}$	$0.78^{0.88}_{0.59}$	$0.61^{0.74}_{0.45}$	$1.39^{1.47}_{1.3}$	$0.19^{0.4}_{0.04}$	$0.44^{0.66}_{0.14}$	$0.27^{0.46}_{0.07}$	$1.52^{1.62}_{1.42}$	$0.36^{0.6}_{0.09}$	$0.5^{0.76}_{0.16}$	$0.38^{0.61}_{0.13}$	$8.76^{9.37}_{8.14}$
SYK	44	4PV0	$0.24^{0.47}_{0.03}$	$0.37^{0.59}_{0.12}$	$0.25^{0.42}_{0.08}$	$1.63^{1.69}_{1.57}$	$0.01^{0.1}_{0.0}$	$-0.17^{0.1}_{-0.41}$	$-0.12^{0.07}_{-0.31}$	$1.69^{1.76}_{1.62}$	$0.02^{0.1}_{0.0}$	$0.04^{0.28}_{-0.2}$	$0.01^{0.17}_{-0.15}$	$10.17^{10.51}_{9.82}$
TNKS2	27	4UI5	$0.16^{0.41}_{0.01}$	$0.41^{0.66}_{0.07}$	$0.29^{0.51}_{0.05}$	$2.2^{2.32}_{2.08}$	$0.14^{0.37}_{0.01}$	$0.32^{0.6}_{-0.02}$	$0.22^{0.45}_{-0.03}$	$1.35^{1.43}_{1.27}$	$0.07^{0.22}_{0.0}$	$0.22^{0.52}_{-0.14}$	$0.14^{0.37}_{-0.1}$	$7.9^{8.41}_{7.4}$
Total	264		$0.43^{0.64}_{0.25}$	$0.64^{0.79}_{0.44}$	$0.49^{0.64}_{0.33}$	$1.68^{1.76}_{1.60}$	$0.09^{0.27}_{0.02}$	$0.21^{0.48}_{-0.08}$	$0.15^{0.36}_{-0.06}$	$1.83^{1.93}_{1.73}$	$0.24^{0.41}_{0.11}$	$0.41^{0.61}_{0.15}$	$0.29^{0.46}_{0.11}$	$8.78^{9.22}_{8.33}$

(A)). Errors in $\Delta\Delta G$ appear to follow a Gaussian distribution (Figure 6 (B)). The data sets contains 12 neutral compounds and 12 positively charged compounds that carry a solvent-exposed basic group. The FEP map contained six perturbations that involved a change in net charge. For five of these transformations, the predicted $\Delta\Delta G_{\text{pred}}$ was within 1.1 kcal/mol of the experimental value (examples are shown in Figure 6 (C)). The remaining charge-changing transformation involved molecule CHEMBL3402762 that had a measured affinity $\text{IC}_{50} < 1$nM (top of the assay, this was set to 1 nM in order to include the compound in the comparison). This compound was predicted to be more potent than CHEMBL3402761 which has a reported $\text{IC}_{50} = 1$ nM. Overall, the perturbations in the c-Met data set demonstrate that transformations involving changes in net charge can be handled reliably by the FEP+ method. Still, the benchmark cases also illustrates the challenges that are involved in predicting charge changes. In the TNKS2 set, the relative affinities within ligands of the same net charge were accurately predicted. The relative affinities between ligands of different net charge, however, displayed larger errors.

For the c-Met case, FEP+ predictions also reproduced the SAR for changing from a carbamate unit to various aromatic heterocycles (Figure 6 (D)). It correctly ranked pyrimidine as more potent than two thiazole variants, imidazole, oxadiazole and pyridazine. It did however not reproduce the increase in potency when going from a pyridine to a pyrimidine

(these two compounds that have binding affinity of IC_{50} = 200 nM and IC_{50} = 40 nM respectively are predicted as 49 nM and 60 nM respectively). This change is still within the typical error limit of the method (1 kcal/mol). We hypothesized that the protonation state of the pyridine may affect the affinity and might explain the difference in potency. But pKa calculations with Jaguar[63] yielded a pKa of 4.5 for the pyridine compound which makes the presence of the charged species highly unlikely.

Comparing FEP+ results on the benchmark with results for Glide docking and Prime MM-GBSA calculations, we found that FEP+ showed significantly better correlation to experimental data and lower or equivalent RMSE for seven of the eight cases (Table 1 and Figure 7). On this benchmark, FEP+ clearly outperformed both simpler methods (Cohen's d for Kendall τ was 1.94 and 1.40 when compared with Glide and Prime respectively indicating a very large effect size[51,52]). In addition, it also clearly outperformed two "null models" (Figure 7 and Table S4; Cohen's d for Kendall τ was 2.25 and 1.94 for comparison with ranking by molecular weight and calculated log P respectively, indicating a huge and very large effect size[51,52]). This result is in qualitative agreement with the data obtained from our in-house projects.

It is important to point out that the benchmark results presented in this study set were obtained in an industry context without extensive optimization following our standard protocol. It may very well be possible to obtain even higher accuracy results with more thorough model optimization. Indeed, for the Eg5 data set, we found that the protein structure–specifically one loop in the vicinity of the ligand–had a large effect on the accuracy of the prediction. We could clearly link this structural change to a set of outliers involving a phenol group that was in contact with this loop (see Figure S4). However, in the context of drug discovery projects, there is usually only limited time for validation and model optimization (typically 2 weeks). In our opinion, the performance reported here gives a realistic view what accuracy can be expected when using FEP+ in an industry setting. We hope this set will be useful to drive further method development in the field. The input struc-

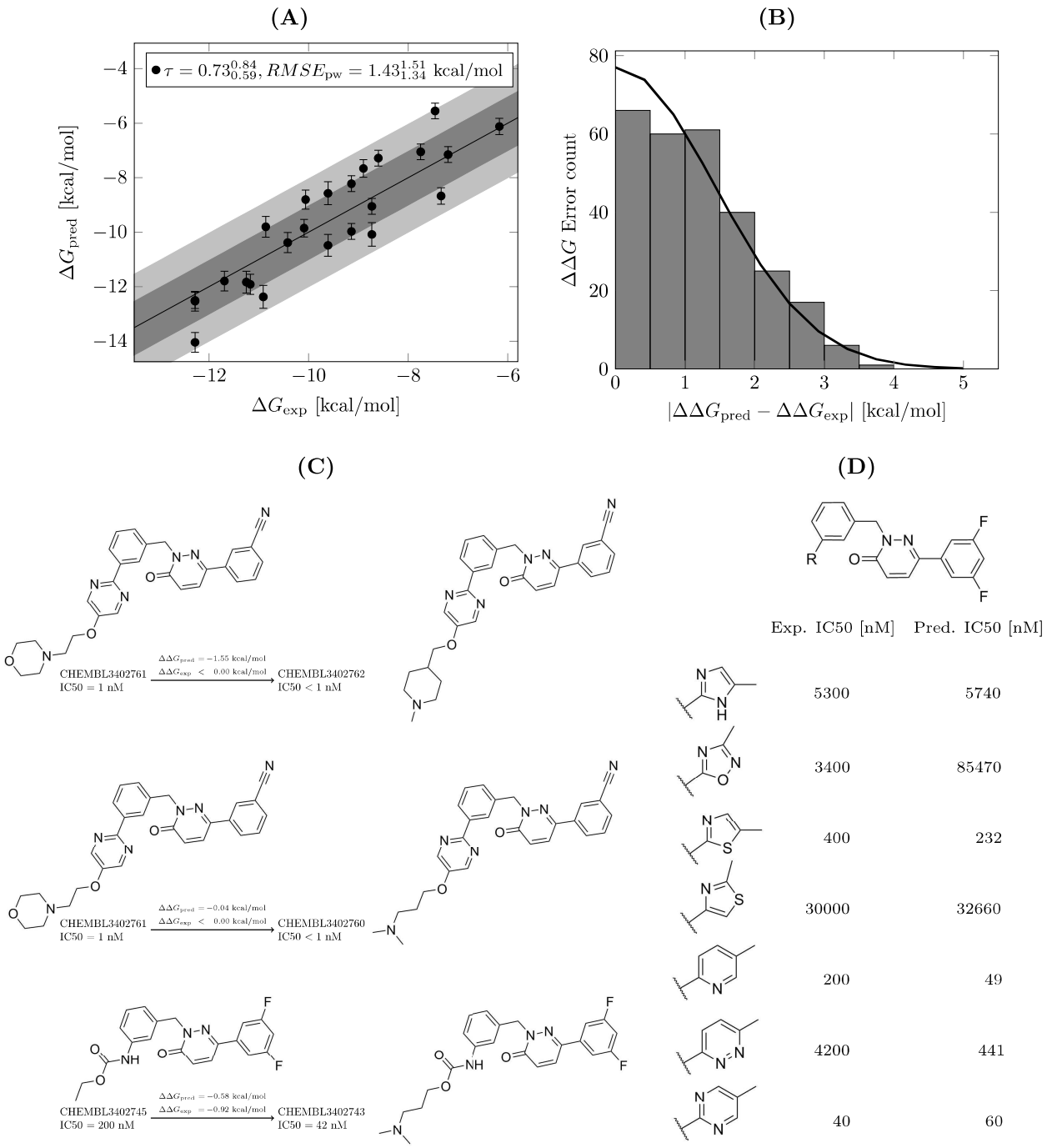

Figure 6 FEP+ results for c-Met benchmark case. (A) Predicted affinities ΔG_{pred} correlate well with experimental affinities ΔG_{exp}. (B) Histogram of the errors for pairwise relative affinities $|\Delta\Delta G_{\mathrm{pred}} - \Delta\Delta G_{\mathrm{exp}}|$. The solid line shows a Gaussian curve with standard deviation 1.43 kcal/mol. (C) Perturbations involving net charge changes in the data set. FEP+ shows good accuracy in predicting these challenging transformations. (D) FEP correctly predicts ranks a series of different aromatic heterocycles substitutions.

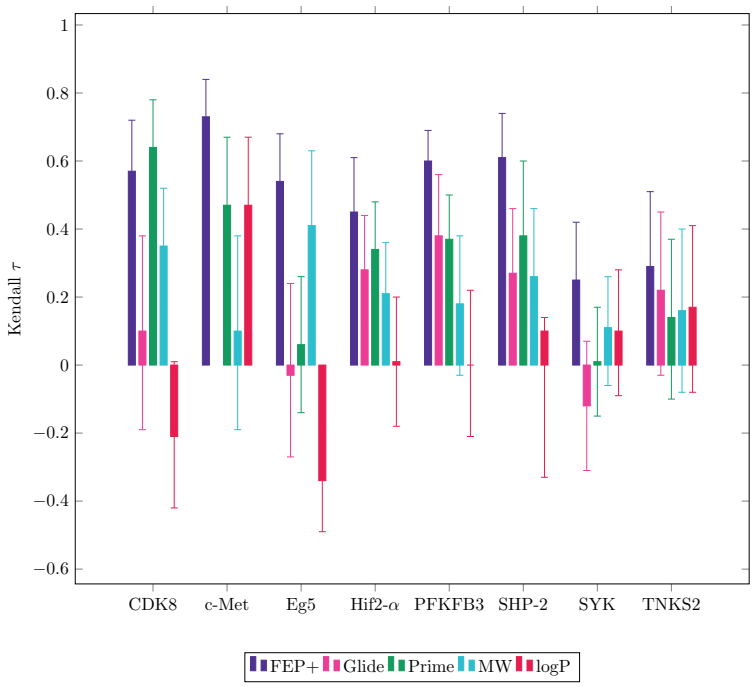

Figure 7 Comparison of Kendall τ for FEP+, Glide, Prime, MW and log P ranking on the eight benchmark cases. Confidence intervals were estimated via bootstrap sampling.

tures for the eight benchmark cases and FEP+ results for 5 ns and 20 ns are available on `www.github.com/MCompChem/fep-benchmark`.

Impact of FEP+ on projects and operational challenges

Despite these overall encouraging prospective FEP+ results and the clear advantage over simpler SBDD methods, throughout the initiative it became clear to us that the usefulness of the prediction and the required level of accuracy to achieve meaningful ranking heavily depend on the the chemical series and the observed dynamic range/affinity distribution in the optimization phase. For example, for target 2/series 1 and target 2/series 3, we found good ranking with FEP+, although the reported RMSE was larger than 1.5 kcal/mol (Table S1 and Figure 3). This was helpful in prioritizing compounds for synthesis and was regarded as valuable by the project chemists. On the other hand, in target 1/series 1-3, we did not obtain any predictive ranking despite having similar accuracy in terms of RMSE. In our experience, lower accuracy can be tolerated in projects that display a larger variation in the

underlying potency distribution of the chemical series. Once optimization has hit an affinity "canyon" in chemical space where small variations do not have a large effect on potency, applying free energy calculations has only limited value for the project (as was the case in target 1/series 1-3 and target 5/series 1). Interestingly, we found that such a situation was often present in (late stage) lead optimization projects. At the start of the initiative, we assumed that lead optimization should be the primary use case for FEP, since the small scale chemical transformations that are typically carried out at this stage are best suited for the method. Yet in contrast to our initial expectation, we noticed that we had better impact on hit-to-lead optimization or on fragment optimization when chemical space was more broadly explored and potency was still a major optimization parameter. Furthermore, at this point, chemistry resources were usually more limited and chemists still faced synthetic challenges that limited the number of molecules that could be synthesized and tested experimentally. In such a situation, prioritizing ideas to focus on the most promising ones was perceived as very valuable and performing these computationally expensive calculations was not regarded as a bottleneck. Additionally, guiding design teams towards compounds that should be made generates a higher impact than advising them which compounds should *not* be made. Assessing the impact on the optimization retrospectively for the latter is not easily done. In terms of domain of applicability, one has to keep in mind that predictivity can be limited in early hit optimization phase, since the chemical transformations encountered in these early stages are typically more challenging for the method.

Strong communication was also crucial for successfully implementing free energy calculations in projects. We experienced how important it is to have a good understanding of the capabilities and limitations of the method in the project team. This helped in selecting compounds for prioritization with FEP+ that were within the domain of applicability. Initially, we faced challenges in using FEP+ in projects because ideas submitted for calculations were outside of the scope of the method. In later stages of the initiative, we focused on ranking custom-built libraries that were by design more amenable for the calculations. A disadvan-

tage of this approach was that in some cases the results of such a library scan were taken up by the project team and used in the design of new molecules, however, this information was combined with other ideas. In this way, it was hard to assess the impact of the method since the exact molecules as predicted by FEP were not synthesized. To avoid this issue, we aimed to predict such molecules later with FEP+ in order to be able to assess the quality of the predictions with our automatic workflow (see above). In general, we found that the acceptance of FEP predictions by chemists was higher when the results could be interpreted or rationalized; e.g., by analyzing interactions or ligand flexibility. We therefore recommend to deliver FEP predictions accompanied by such an analysis, especially focusing on those compounds that were predicted to be the best and the worst binders.

In summary, to make best use of free energy calculations in projects, prediction accuracy, domain of applicability, key optimization challenges, synthetic accessibility of the chemical series of interest and timing in project have to be carefully balanced. We found that screening large custom-built libraries – ideally designed with non-potency optimization parameters in mind – to be an effective way of providing added value. For these libraries, we screen at least 50-100 ideas. We aim to screen 5-10 times more ideas than the maximum number of compounds that can be selected for synthesis (this is in line with a recent publication[64]).

Conclusion

Free energy calculations are more and more frequently used in pharmaceutical industry and have become a powerful addition in the computational chemist's toolbox. Here, we present data from using FEP+ prospectively in a large number of in-house drug discovery projects. We obtained good accuracy for a large variety of targets and chemical series. Yet, the number of targets where we were able to obtain high accuracy predictions (RMSE $<$ 1 kcal/mol) was limited. These results were in line with those obtained on a new, public benchmark set. In real world drug design projects, structural information and availability of ligand series

with large affinity spread is not always given. In addition to the accuracy of the prediction, we identified multiple important operational factors that affect the impact of the method in projects. In the near future, we envision FEP+ as an expert tool to support FEP-enabled projects with large-scale library calculations.

Experimental

Protein structure preparation

Protein structures were downloaded from the PDB (www.rcsb.org) or from our in-house database and imported into Maestro.[65] If multiple chains were present and there was no indication that the multimer was the biologically relevant form, structures were split and each chain was processed separately. Prime Homology modeling[66,67] was used to remove mutations introduced in the crystallization constructs (if any), build missing side-chains and missing loops. If loops had been added, these were subsequently refined with the Prime Loop Refinement protocol. Sequence alignment was performed with ClustalW and adapted manually if necessary. The homology model was built using the knowledge-based approach and taking the inhibitor into account. For loop refinement, recommended settings depending on the length of the loop were used. If larger domains (more than 20 residues) were missing, these parts of the structure were not added. In case of multiple structures for a given target, loop structure were also modeled based on alternative structures. The structures were then processed with the Protein Preparation Wizard in Maestro.[65,68] Hydrogens were added, all crystal waters were retained and the termini were capped. Co-crystal ligand protonation states were evaluated with Epik.[69] Protein protonation states were determined with PropKa[70,71] and hydrogen bonds were optimized. The structures were finally minimized with an RMSD cutoff of 0.3 Å (default settings).

Ligand structure preparation

Corina was used to generate 3D coordinates based on the 2D representation. Structures were then processed with LigPrep enumerating possible stereoisomers and assigning protonation states with Epik.[69,72] For positioning the ligands into the binding site, we either used the Flexible Ligand Alignment tool or Glide core-constrained docking[73] based on a reference structure (in most cases, the X-ray ligand). The core was defined by maximum common substructure or by a custom SMARTS pattern for molecules involving core changes. We ran Glide core-constrained docking using standard precision settings. Structures were analyzed with the Force Field Builder to detect missing torsion parameters. If necessary, new custom parameters were generated by the Force Field Builder protocol.

Prospective free energy calculations in in-house projects

Prospective free energy calculations were performed using the Schrödinger FEP+ method[5] with Schrödinger suite versions 2016-4 to 2018-3. Up to version 2018-1, the OPLS3 force field[12] was used with custom parameters generated by the Force Field Builder. For calculations performed with version 2018-2 and higher, the OPLS3e force field[14] with custom parameters was used. Prospective calculations were run with optimal settings as determined by the validation study. In most cases, these corresponded to the default settings, in some cases, the sampling was extended beyond 5 ns per window. Calculations were analyzed using the FEP+ GUI.[65] If convergence issues were found, simulations were extended if time and computing resources allowed. In all cases, unconverged edges and edges with high hysteresis values were removed from the final map. Finally, all predicted ΔG values > -5.81 kcal/mol were set to -5.81 kcal/mol to allow better comparison to experiment (typical value for bottom of the assay).

In-house in vitro assays

For all targets except target 8 and target 2, biochemical enzyme activity assays were used to determine the binding affinity of the molecules. For target 8, a mixture of functional and ITC measurements was used for comparison with FEP+ results (molecules that showed no activitiy in the functional assay were not profiled with ITC). For target 2, affinities were determined using a cellular phenotypic assay.

Free energy calculations on benchmark data set

The full data set is summarized in Table 1. Protein co-crystal structures were downloaded from the Protein Data Bank (www.rcsb.org). Ligand structures and affinities (IC_{50}) were extracted manually from papers and/or patents.[53–62] IC_{50} values were converted into free energy values using the equation $\Delta G \approx k_B T \log \mathrm{IC}_{50}$. Protein structures and ligand structures were prepared for free energy calculations as described above. For the benchmark, free energy calculations with FEP+ were run using Schrodinger suite version 2018-3 with the OPLS3e force field.[14] Prepared structures were loaded into the FEP+ panel and affinity data were added. Maps were generated with default settings (optimal topology). No further modifications were made to the map. FEP+ jobs were run for 5 ns and 20 ns sampling time per λ-window. The output was analyzed with the FEP+ panel in Maestro. In contrast to the prospective calculations, we did not modify the final map – e.g., remove edges with high hysteresis – to allow for better reproducibility of the results.

Comparison to other methods

For comparison, we docked ligands into their respective protein structure using the Glide standard precision ligand docking workflow[73] in Schrödinger suite 2018-3 with default settings. As receptor, the protein structure used for the free energy calculations was used (all water molecules were deleted). The ligands were ranked according to Glide gscore and this

score was compared to experimental affinities. MM-GBSA calculations were performed with the Prime MM-GBSA module in Schrödinger suite 2018-3 using default settings. The protein structure without water molecules was used as receptor and the ligand poses were taken from the FEP+ input. The ligands were ranked according to "MMGBSA dG Bind" score. Molecular weight and logP were calculated for the neutral ligands using Maestro.[65]

Analysis

Convergence, hysteresis and interaction analysis were performed using the FEP+ GUI in Maestro. Correlations statistics and errors were evaluated using Python numpy, scipy[74] and bootstrapped libraries. 90% symmetric confidence intervals (90% CI) for all performance metrics were calculated using bootstrap by resampling all data sets with replacement, with 10000 resampling events. CIs were estimated for all performance metrics and reported as $x_{x_{\text{low}}}^{x_{\text{high}}}$ where x is the mean statistic calculated from the complete data set (e.g., RMSE), and x_{low} and x_{high} are the values of the statistic at the 5th and 95th percentiles of the value-sorted list of the bootstrap samples. Data curation and extraction of experimental data for prospective FEP calculations from our in-house database were carried out using in-house custom scripts. IC_{50} or K_d values were converted to free energies using the equations

$$\Delta G_{\text{exp}} \approx k_B T \log \text{IC}_{50}$$

$$\Delta G_{\text{exp}} = k_B T \log K_d.$$

Acknowledgement

The authors thank Vanita Sood, Klaus Urbahns, Thomas Fürst, Stefan Oschmann and the Digitizing Merck initiative for funding and mentoring and Jörg Weiser for support with the Schrödinger software and licenses. They also thank Lukas Friedrich for critically reading the manuscript.

Supporting Information Available

The following files are available free of charge.

- supplementary.pdf: Supplementary Figures and Tables.
- tables.zip: Tables and Supplementary tables as csv files.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Jiang, W.; Roux, B. Free energy perturbation Hamiltonian replica-exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations. *J. Chem. Theory Comput.* **2010**, *6*, 2559–2565.

(2) Wang, L.; Friesner, R. A.; Berne, B. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B* **2011**, *115*, 9431–9438.

(3) Gallicchio, E.; Levy, R. M. Advances in all atom sampling methods for modeling protein–ligand binding affinities. *Curr. Opin. Struct. Biol.* **2011**, *21*, 161–166.

(4) Kaus, J. W.; McCammon, J. A. Enhanced ligand sampling for relative protein–ligand binding free energy calculations. *J. Phys. Chem. B* **2015**, *119*, 6190–6197.

(5) Wang, L. 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**, *137*, 2695–2703.

(6) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(7) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell Jr., A. D. 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*, 671–690.

(8) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.

(9) Knight, J. L.; Yesselman, J. D.; Brooks III, C. L. Assessing the quality of absolute hydration free energies among CHARMM-compatible ligand parameterization schemes. *J. Comput. Chem.* **2013**, *34*, 893–903.

(10) Martins, S. A.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Prediction of solvation free energies with thermodynamic integration using the general amber force field. *J. Chem. Theory Comput.* **2014**, *10*, 3570–3577.

(11) Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **2017**, *45*, W331–W336.

(12) Harder, E. et al. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* **2015**, *12*, 281–296.

(13) Nerenberg, P. S.; Head-Gordon, T. New developments in force fields for biomolecular simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 129–138.

(14) Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1863–1874.

(15) Träg, J.; Zahn, D. Improved GAFF2 parameters for fluorinated alkanes and mixed hydro-and fluorocarbons. *J. Mol. Model.* **2019**, *25*, 39.

(16) Tanner, D. E.; Phillips, J. C.; Schulten, K. GPU/CPU algorithm for generalized Born/solvent-accessible surface area implicit solvent calculations. *J. Chem. Theory Comput.* **2012**, *8*, 2521–2530.

(17) Salomon-Ferrer, R.; GoÌĹtz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.

(18) Bergdorf, M.; Baxter, S.; Rendleman, C. A.; Shaw, D. E. Desmond/GPU Performance as of October 2015. *DE Shaw Research Technical Report DESRES/TR–2015* **2015**, *1*.

(19) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(20) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comp. Biol.* **2017**, *13*, e1005659.

(21) Mermelstein, D. J.; Lin, C.; Nelson, G.; Kretsch, R.; McCammon, J. A.; Walker, R. C. Fast and flexible gpu accelerated binding free energy calculations within the amber molecular dynamics package. *J. Comput. Chem.* **2018**, *39*, 1354–1358.

(22) Lee, T.-S.; Cerutti, D. S.; Mermelstein, D.; Lin, C.; LeGrand, S.; Giese, T. J.; Roitberg, A.; Case, D. A.; Walker, R. C.; York, D. M. GPU-accelerated molecular dynamics and free energy methods in Amber18: performance enhancements and new features. *J. Chem. Inf. Model.* **2018**, *58*, 2043–2050.

(23) Kutzner, C.; Páll, S.; Fechner, M.; Esztermann, A.; de Groot, B. L.; Grubmüller, H. More bang for your buck: Improved use of GPU nodes for GROMACS 2018. *J. Comp. Chem.* **2019**, *40*, 2418–2431.

(24) Gong, X.; Chiricotto, M.; Liu, X.; Nordquist, E.; Feig, M.; Brooks III, C. L.; Chen, J. Accelerating the Generalized Born with Molecular Volume and Solvent Accessible Surface Area Implicit Solvent Model Using Graphics Processing Units. *J. Comp. Chem.* **2019**,

(25) Vanommeslaeghe, K.; MacKerell Jr, A. D. Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144–3154.

(26) Loeffler, H. H.; Michel, J.; Woods, C. FESetup: Automating Setup for Alchemical Free Energy Simulations. *J. Chem. Inf. Model.* **2015**, *55*, 2485–2490, PMID: 26544598.

(27) Klimovich, P. V.; Mobley, D. L. A Python tool to set up relative free energy calculations in GROMACS. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 1007–1014.

(28) Liu, S.; Wu, Y.; Lin, T.; Abel, R.; Redmann, J. P.; Summa, C. M.; Jaber, V. R.; Lim, N. M.; Mobley, D. L. Lead optimization mapper: automating free energy calculations for lead optimization. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 755–770.

(29) Wang, L.; Deng, Y.; Wu, Y.; Kim, B.; LeBard, D. N.; Wandschneider, D.; Beachy, M.; Friesner, R. A.; Abel, R. Accurate modeling of scaffold hopping transformations in drug discovery. *J. Chem. Theory Comput.* **2016**, *13*, 42–54.

(30) Chen, W.; Deng, Y.; Russell, E.; Wu, Y.; Abel, R.; Wang, L. Accurate calculation of relative binding free energies between ligands with different net charges. *J. Chem. Theory Comput.* **2018**, *14*, 6346–6358.

(31) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)–Round XII. *Proteins: Struct., Funct., Bioinf.* **2018**, *86*, 7–15.

(32) Lensink, M. F.; Velankar, S.; Wodak, S. J. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins: Struct., Funct., Bioinf.* **2017**, *85*, 359–377.

(33) Rizzi, A.; Murkli, S.; McNeill, J. N.; Yao, W.; Sullivan, M.; Gilson, M. K.; Chiu, M. W.; Isaacs, L.; Gibb, B. C.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6 host–guest binding affinity prediction challenge. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 937–963.

(34) Gaieb, Z.; Parks, C. D.; Chiu, M.; Yang, H.; Shao, C.; Walters, W. P.; Lambert, M. H.; Nevins, N.; Bembenek, S. D.; Ameriks, M. K.; Mirzadegan, T.; Burley, S. K.; Amaro, R. E.; Gilson, M. K. D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *J. Comput.-Aided Mol. Des.* **2019**, *33*, 1–18.

(35) Genheden, S. Are homology models sufficiently good for free-energy simulations? *J. Chem. Inf. Model.* **2012**, *52*, 3013–3021.

(36) Park, H.; Lee, S. Homology modeling, force field design, and free energy simulation studies to optimize the activities of histone deacetylase inhibitors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 375–388.

(37) Cappel, D.; Hall, M. L.; Lenselink, E. B.; Beuming, T.; Qi, J.; Bradner, J.; Sherman, W. Relative binding free energy calculations applied to protein homology models. *J. Chem. Inf. Model.* **2016**, *56*, 2388–2400.

(38) Brown, S. P.; Muchmore, S. W.; Hajduk, P. J. Healthy skepticism: assessing realistic model performance. *Drug Discovery Today* **2009**, *14*, 420–427.

(39) Shirts, M. R.; Mobley, D. L.; Brown, S. P. In *Drug Design: Structure-and Ligand-Based Approaches*; Merz Jr., K. M., Ringe, D., Reynolds, C. H., Eds.; Cambridge University Press: 32 Avenue of the Americas, New York, USA, 2010; pp 61–86.

(40) de Oliveira, C.; Yu, H. S.; Chen, W.; Abel, R.; Wang, L. Rigorous Free Energy Perturbation Approach to Estimating Relative Binding Affinities between Ligands with Multiple Protonation and Tautomeric States. *J. Chem. Theory Comput.* **2018**, *15*, 424–435.

(41) Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Protonation changes upon ligand binding to trypsin and thrombin: structural interpretation based on pKa calculations and ITC experiments. *J. Mol. Biol.* **2007**, *367*, 1347–1356.

(42) Petukh, M.; Stefl, S.; Alexov, E. The role of protonation states in ligand-receptor recognition and binding. *Curr. Pharm. Des.* **2013**, *19*, 4182–4190.

(43) Martin, Y. C. Let's not forget tautomers. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 693.

(44) Mongan, J.; Case, D. A. Biomolecular simulations at constant pH. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157–163.

(45) Stern, H. A. Molecular simulation with variable protonation states at constant pH. *J. Chem. Phys.* **2007**, *126*, 04B627.

(46) Wallace, J. A.; Shen, J. K. Continuous constant pH molecular dynamics in explicit solvent with pH-based replica exchange. *J. Chem. Theory Comput.* **2011**, *7*, 2617–2629.

(47) Swails, J. M.; York, D. M.; Roitberg, A. E. Constant pH replica exchange molecular dynamics in explicit solvent using discrete protonation states: implementation, testing, and validation. *J. Chem. Theory Comput.* **2014**, *10*, 1341–1352.

(48) Arthur, E. J.; Brooks III, C. L. Efficient implementation of constant pH molecular dynamics on modern graphics processors. *J. Comp. Chem.* **2016**, *37*, 2171–2180.

(49) Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping atom types in force fields using direct chemical perception. *J. Chem. Theory Comput.* **2018**, *14*, 6076–6092.

(50) Abel, R.; Wang, L.; Mobley, D. L.; Friesner, R. A. A critical review of validation, blind testing, and real-world use of alchemical protein-ligand binding free energy calculations. *Curr. Top. Med. Chem.* **2017**, *17*, 2577–2585.

(51) Cohen, J. *Statistical power analysis for the behavioral sciences*; Lawrence Earlbaum Associcates, 1988.

(52) Sawilowsky, S. S. New effect size rules of thumb. *J. Mod. Appl. Stat. Methods* **2009**, *8*, 26.

(53) Schiemann, K.; Mallinger, A.; Wienke, D.; Esdar, C.; Poeschke, O.; Busch, M.; Rohdich, F.; Eccles, S. A.; Schneider, R.; Raynaud, F. I.; Czwodrowski, P.; Musil, D.; Schwarz, D.; Urbahns, K.; Blagg, J. Discovery of potent and selective CDK8 inhibitors from an HSP90 pharmacophore. *Bioorg. Med. Chem. Lett.* **2016**, *26*, 1443–1451.

(54) Dorsch, D.; Schadt, O.; Stieber, F.; Meyring, M.; Grädler, U.; Bladt, F.; Friese-Hamim, M.; Knuehl, C.; Pehl, U.; Blaukat, A. Identification and optimization of pyridazinones as potent and selective c-Met kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 1597–1602.

(55) Schiemann, K.; Finsinger, D.; Zenke, F.; Amendt, C.; Knöchel, T.; Bruge, D.; Buchstaller, H.-P.; Emde, U.; Stähle, W.; Anzali, S. The discovery and optimization of hexahydro-2H-pyrano [3, 2-c] quinolines (HHPQs) as potent and selective inhibitors of the mitotic kinesin-5. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 1491–1495.

(56) Dixon, D. D.; Grina, J.; Josey, J. A.; Rizzi, J. P.; Schlachter, S. T.; Wallace, E. M.;

Wang, B.; Wehn, P.; Xu, R.; Yang, H. Aryl ethers and uses thereof. 2018; US Patent 9,896,418.

(57) Wehn, P. M. et al. Design and Activity of Specific Hypoxia-Inducible Factor-2α (HIF-2α) Inhibitors for the Treatment of Clear Cell Renal Cell Carcinoma: Discovery of Clinical Candidate (S)-3-((2, 2-Difluoro-1-hydroxy-7-(methylsulfonyl)-2, 3-dihydro-1 H-inden-4-yl) oxy)-5-fluorobenzonitrile (PT2385). *J. Med. Chem.* **2018**, *61*, 9691–9721.

(58) Boutard, N. et al. Discovery and Structure–Activity Relationships of N-Aryl 6-Aminoquinoxalines as Potent PFKFB3 Kinase Inhibitors. *ChemMedChem* **2019**, *14*, 169–181.

(59) Garcia Fortanet, J. et al. Allosteric inhibition of SHP2: identification of a potent, selective, and orally efficacious phosphatase inhibitor. *J. Med. Chem.* **2016**, *59*, 7773–7782.

(60) Chen, Y.-N. P. et al. Allosteric inhibition of SHP2 phosphatase inhibits cancers driven by receptor tyrosine kinases. *Nature* **2016**, *535*, 148.

(61) Currie, K. S. et al. Discovery of GS-9973, a selective and orally efficacious inhibitor of spleen tyrosine kinase. *J. Med. Chem.* **2014**, *57*, 3856–3873.

(62) Buchstaller, H.-P.; Anlauf, U.; Dorsch, D.; Kuhn, D.; Lehmann, M.; Leuthner, B.; Musil, D.; Radtki, D.; Ritzert, C.; Rohdich, F.; Schneider, R.; Esdar, C. Discovery and Optimization of 2-Arylquinazolin-4-ones into a Potent and Selective Tankyrase Inhibitor Modulating Wnt Pathway Activity. *J. Med. Chem.* **2019**,

(63) Bochevarov, A. D.; Harder, E.; Hughes, T. F.; Greenwood, J. R.; Braden, D. A.; Philipp, D. M.; Rinaldo, D.; Halls, M. D.; Zhang, J.; Friesner, R. A. Jaguar: A high-performance quantum chemistry software program with strengths in life and materials sciences. *Int. J. Quantum Chem.* **2013**, *113*, 2110–2142.

(64) Abel, R.; Manas, E. S.; Friesner, R. A.; Farid, R. S.; Wang, L. Modeling the value of predictive affinity scoring in preclinical drug discovery. *Curr. Opin. Struct. Biol.* **2018**, *52*, 103–110.

(65) Schrödinger Release 2018-3: Maestro, Schrödinger, LLC, New York, NY, 2018.

(66) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 351–367.

(67) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002**, *320*, 597–608.

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

(69) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.

(70) Søndergaard, C. R.; Olsson, M. H.; Rostkowski, M.; Jensen, J. H. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values. *J. Chem. Theory Comput.* **2011**, *7*, 2284–2295.

(71) Olsson, M. H.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

(72) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the compre-

hensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 591–604.

(73) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(74) Virtanen, P. et al. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv preprint* **2019**, arXiv:1907.10121.

Graphical TOC Entry