INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Applications of Modern pharmacological research uses artificial intelligence in alchemical free energy calculations

[1]Ramesh Kumar C, Research Scholar, Department of Chemistry , Radha Govind University, Ramgarh,Jharkhand.

[2]Dr.Shiv Brat Singh ,Assistant Professor , Department of Chemistry ,Radha Govind University, Ramgarh, Jharkhand

Abstract-The present thesis delves into the integration of alchemical free energy approaches (AFE) with machine-learning (ML) techniques to augment computer-aided drug discovery (CADD). The primary focus of the work is to analyze the potential synergies that can be attained between various components. The present study postulates that while individual machine learning (ML) and augmented feature engineering (AFE) techniques in computer-aided drug design (CADD) hold considerable promise, their integration in a way that optimizes the strengths of each component could yield added benefits.

The utilization of physics-based AFE calculations has emerged as a highly accurate and precise sub-kcal·mol−1 approach for forecasting the binding affinities of ligand-protein interactions. This has resulted in their widespread adoption in supporting drug design workflow projects. The rapid development of data-driven machine learning approaches can be attributed to the exponential expansion in computer hardware capabilities. However, it is worth noting that these approaches still exhibit lower accuracy levels compared to experimental binding affinities when it comes to drug discovery. The former approach employs statistical mechanics, whereas the latter involves signal interpolation from extensive training data regions. The thesis will commence with a historical and theoretical introduction to drug development, AFE calculations, and ML approaches. Subsequently, it will present several investigations that substantiate the aforementioned hypothesis at various stages in the AFE process.

Initially, precise values of hydration free energies are computed through the employment of AFE and ML methodologies. After being trained on a section of the FreeSolv database, the hybrid AFE/ML technique demonstrated superior performance compared to the majority of SAMPL4 submissions. It seems that the utilization of AFE/ML may offer certain benefits over conducting independent AFE computations, such as the possibility of necessitating a reduced training set size. Correction terms derived from machine learning can also be applied to related AFE simulation methods. This approach has the potential to efficiently improve AFE calculations and pinpoint specific compounds that would derive the greatest advantage from tailored force field parameterization.

Furthermore, research has been initiated on the generation of AFE networks via data analysis. Practitioners must exercise caution while conducting AFE calculations by giving due consideration to the amalgamation of alchemical transformations among ligands in congeneric series. AFE networks refer to a set of edges that are associated with AFE computations for all ligands, also known as nodes. It seems that there could be some issues related to network configurations that may pose challenges during the AFE setup phase. These challenges could potentially affect the scalability and transferability of the AFE software across various platforms.

The methodology is dependent on the reliability of AFE transformation estimates provided by experts, despite the automated construction of the AFE network. The present document presents a novel data-driven approach based on a graph siamese neural network, which serves as an alternative to the current state-of-the-art methods. The AFE ML study employs RBFE-Space, a dataset that has been recently developed. The methodology employed in this thesis showcases noteworthy enhancements in the performance of AFE network generation, through the utilization of cutting-edge techniques. The RBFE-Space platform is a versatile and open-source solution that enables seamless integration with other AFE applications. This facilitates the transfer of the network generator, enhancing overall efficiency. The deep learning model represents the initial reliable machine learning predictor of AFE transformation reliabilities.

The reliability of AFE computations decreases for transformations involving more than 5 heavy atoms. This study examines the efficacy of performing individual transformations of running charge, Van der Waals, and bond parameters with variable allocation per step, as opposed to the standard practice of transforming all parameters in a single step in most AFE workflows. The MultiStep protocol is more advantageous for the bound leg as compared to the one-step ("SoftCore") approach, whereas the free leg does not exhibit any such benefits. Based on Cresset's additional research, it has been determined that the Softcore methodology and the MultiStep technique offer comparable benefits. This study emphasizes the advantages of analyzing a First Episode Psychosis (FEP) approach and comparing it with an alternative strategy.

1.Introduction

The introductory section of this thesis aims to provide the reader with the necessary foundational information to effectively analyze the research presented in the following chapters. Despite making an attempt to incorporate a substantial amount of theoretical backing within the constraints of conciseness, it is recommended to pursue additional literature, particularly when it pertains to the subject matter at hand.

The introduction is organized in a top-down manner, beginning with an overview of the pharmaceutical drug discovery pipeline and the significance of computer-aided drug design in this process. This is followed by a discussion of the fundamental theoretical principles underlying the binding of ligands to proteins. Subsequently, a conclusion will be presented to provide a comprehensive overview of the outcomes derived from the research conducted in the thesis. Subsequently, a comprehensive introduction to molecular simulation is provided, succeeded by an analysis of its correlation with alchemical free energy calculations. In summary, the present study offers an exposition of the theoretical underpinnings of machine learning, as applied to the domain of computer-assisted drug design.

A. The contemporary pharmaceutical research and development environment

Finding new ways to treat diseases is essential to expanding the field of global healthcare and getting closer to the ultimate aim of eradicating sickness altogether. An illness's phenotype may be improved by administering a medicine that interacts with a therapeutic target in a way that alters the target's biological function, which is often the main goal of a drug discovery effort. The process of developing new medicines is lengthy and expensive. As a consequence of massive international efforts to identify new medications, several different classes of medicinal agents have been produced. Small-molecule inhibitors, monoclonal antibodies, and vaccines are all examples of this kind of treatment. The research is being conducted in both a university setting and a business one.3 It is important to highlight that the next parts of this introduction will not go into other categories of therapeutic agents, since the focus of this thesis is restricted to the research of small-molecule medications. Drug discovery will hereafter be used to refer to the search for small-molecule medicines.

Drug development is notoriously expensive, both monetarily and in terms of time invested. Some estimates put the cost of developing a new medication at above £1 billion. Current estimates place that figure between £0.3 billion and £0.8 billion. Researching a new drug may take anywhere from ten to fifteen years, which is a big issue that might hold down pharmaceutical campaigns. The aforementioned span of time begins with the launch of a preclinical program and ends when a pharmaceutical product is ready for widespread commercial distribution. The estimates made above do not take into consideration basic research into the processes causing the illness phenotype. Many longitudinal studies span many decades and include numerous participants, most of whom work in academia. The costs associated with these methods are difficult to predict.

Multiple variables contribute to the difficulty of this procedure. The meat of the issue is that it's difficult to create a drug with desirable effects and minimal side effects (in terms of pharmacokinetics, dynamics, and toxicity, among other things). If a medicine has the potential for mass production and has the desired therapeutic effect, we may say that it is effective. Recent systematic estimates suggest a potential rate of up to 86.2% across a sample of 5764 unique pharmaceutical research firms, demonstrating the prevalence of failure is high (and according to the particular characteristics of the sickness).

B. The current status of drug development and research in the pharmaceutical industry

The traditional model of the pharmaceutical pipeline (refer to figure 1.1) consists of a series of discrete steps that function as a system, taking in molecular candidates and producing marketable drugs. In the early phases of drug development, a large number of molecular candidates, typically between 103 and 109, are screened using experimental and computational methods to find a lead chemical. A chemical entity that has achieved the desired pharmacological or biological effects but requires additional modifications in its structural composition to increase its binding affinity towards the therapeutic target or to improve its metabolic and toxicological profiles is considered a lead compound in the scientific community. The first step in the drug discovery process is to zero in on a viable therapeutic target, after which the creation of a lead molecule comes into play. After the lead drug has been identified, it is subjected to in vivo animal experiments to assess its physiologic response prior to human studies, which are known as pre-clinical trials. Next, three consecutive clinical studies are performed to evaluate (1) safety, (2) indication effectiveness, and (3) population-level efficacy. After these tests are complete, the medicine may be submitted for regulatory clearance and put on the market. The estimated rates of attrition in the first, second, and third stages of a clinical study are 86.2%, 79%, and 41%, respectively. The bulk of the expenditures associated with drug discovery are tied to the massive infrastructures of experimentation required for Phases 1-3 of clinical trials. For this reason, it's crucial that clinical applicants be of a high calibre to increase their chances of advancing through the various stages of the clinical trial. For this reason, it is clear that methods that help drug researchers save time and money while still producing higher-quality drug candidates may have a significant effect on the discovery of new therapeutics.

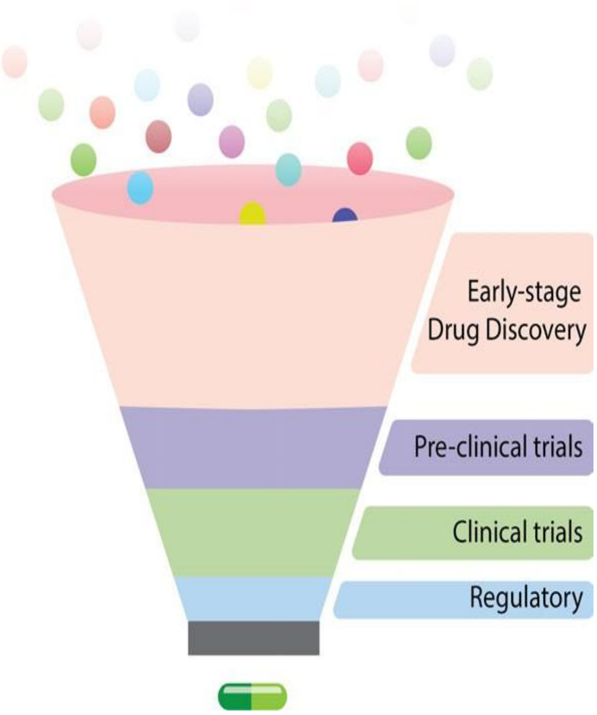

Fig.1: This funnel-like diagram represents the drug discovery pipeline, a frequent metaphor for depicting the whole process of developing a medicine for commercial release. Each coloured circle denotes a potential new medicine, with only one drug making it to market at the end of the funnel.

C. The present study involves the computation of relative binding free energy

Free energy calculations using molecular dynamics (MD) simulations are referred to as alchemical free energy (AFE) computations. Alchemy is a frequent term for the practice of attempting to achieve results via physical rather than chemical means. These equations are used to determine the Gibbs free energy shift for a certain (alchemical) process with NPT apparatuses. Estimating the free energy involved in the binding of tiny molecules to a membrane and computing the change in free energy due to a conformational shift that involves overcoming a high energy barrier are both possible using the AFE calculations. The free energy change associated with a binding mutation in a protein residue may also be calculated in this way. In what follows, we'll discuss how to calculate the RBFE, or relative binding free energy. In this introductory chapter, we use the phrase "RBFE calculations," sometimes known as "de-calculations."

The rate of disso- ciation, which can be on the microsecond timescale even for millimolar binders97 and reaches the microsecond to second timescale for a typical drug103,104, typically dominates the computational cost of these calculations, which have been used to estimate binding affinities96,97 or to gain insights into the binding pathways and kinetics of receptor-ligand systems98-102. Software programs for molecular dynamics vary in how efficiently they do computations based on system size and other simulation parameters. These bundles may accomplish hundreds of ns/day with the help of high-end GPUs. However, this computational performance is neither attractive nor relevant, and hence is impractical for use in pharmaceutical drug development. The binding free energy may also be determined by generating potential of mean force profiles along a reaction coordinate, which is an alternative method. These methods, however, need knowledge of a high-probability binding route in advance, which is not always easy to come by, especially in the prospective settings that are so common in drug development.

A decade into the 21st century, after years of development beginning in the 1980s, RBFE calculations became the first widely used method for reliably predicting ligand binding affinities at a high enough level to support hit-to-lead and lead-optimization campaigns in medicinal chemistry. 111 Since then, the field has advanced to the point where extensive calculations (of the order of hundreds of compounds) can be run in the span of only a few days (given sufficient hardware), allowing computational chemists to provide medicinal chemists with accurate predictions in support of SAR studies at a much faster pace than synthesising each compound individually.

Initial Steps Alchemical transformations are used in the RBFE simulations to simulate ligand interactions. For a particular protein target with many ligands, this entails picking a collection of ligand pairings. These sets may be very small (five to fifteen) for benchmarking reasons, or they may be rather big (fifteen to one hundred) for lead-optimization initiatives. The RBFE software packages have a number of different methods for recommending the transformations to be calculated (see Figure 1.4A).113,114 As certain transformations are more likely to be trustworthy than others, perturbation networks are often used to display the collection of edges suggested for a series of ligands, allowing users to include/exclude transfor- mations depending on user experience. For large-scale projects with several ligands, a star-shaped network is often used, with the reference ligand being the lead molecule currently undergoing optimization. Figure 1.4B shows that at this point in the process, the force field assignment is performed on both the ligands and proteins. If you need more information, look at Paragraph 1.4.2.

This is the manufacturing step. The transformation is often divided into numerous bins using a decoupling parameter, which is based on the definition of molecular transformation among the members of each ligand transformation. Subsequently, the parameters are modified in a bin-wise method, with more disturbed parameters being stored in each bin. Each intermediate system consists of atomistic parameters that have been gradually changed from the endpoints, which include the atomistic parameters of both ligand ends. Using a specific molecular dynamics engine, these windows are simulated one at a time (figure 1.4C), a procedure that often accounts for the bulk of the walltime needed for RBFE simulations. The 1.5.1

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

step simulates both the tied and unbound legs. To further facilitate real-time investigation of hysteresis in both directions, it is preferable to do bidirectional edge simulations that include both A B and B A. By bringing edge hystereses closer to 0 kcalmol1, certain RBFE implementations modify Gbound predictions. Several methods exist for graphically representing the atomic change between two ligand terminals. Single and dual techniques are particularly common. The former includes as little indirect conversion as feasible, whereas the latter involves just conversion to and from non-interacting dummy atoms. Computing procedures associated with topology.

In the scientific literature, this phenomenon is referred to as binding free energy or relative free energy perturbation (FEP).

The phase in which an estimate of free energy is made. After simulations have been run, an estimate of the relative free energy across the decoupling parameter is performed using estimators such as Thermodynamic Integration (TI) or the more modern Multistate Bennett Acceptance Ratio (MBAR) for the perturbation in both the bound and free phase. These estimators are used to calculate the relative free energy across the decoupling parameter. need a theoretical framework that encompasses this estimate in a more thorough manner. At this point in the RBFE program, estimates of the pairwise relative free energy of binding (Gbound) are being generated for all of the proposed changes.

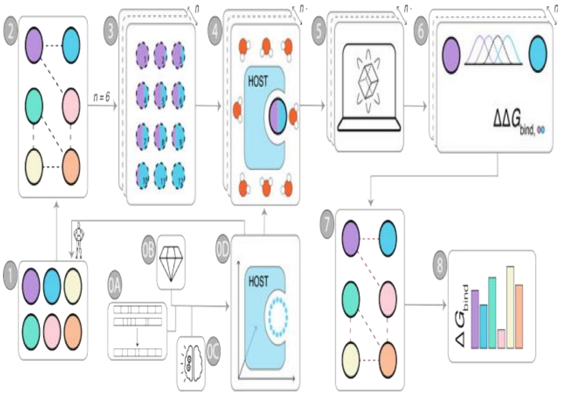

Fig.2 :This research describes RBFE calculating campaign procedure. Homology modelling (A), crystallography (B), and machine learning (C) provide three-dimensional protein structure references. Docking programs manually fill a protein structure's binding pocket with n ligands. A perturbation network determines the series' ligand perturbations. Each edge's -windows are transformed. The reference protein and ligand transformations are solved together. GPUs run simulations. The relative free energy of binding during transformation may be estimated from the -window simulations. After finishing the perturbation network edges, the pairwise -Gbind values are analyzed to determine each ligand's compared to a reference. Gbind predictions per ligand may be used to benchmark experimental results or drive lead optimization.

The stage that deals with analysis

The usage of the initial perturbation network is what is used to accomplish the task of determining the Gbound values for each ligand. In this procedure, it is common practice to make modifications for cycle closures. This is because, in line with the law of conservation of energy, it is anticipated that a cycle of ligands' free energies will have a net energy of 0 kcalmol1 at the end of the cycle. The execution of calculations for both orientations of an edge and then changing the forward and backward free energy estimations associated with the cycle in order to obtain a net energy of 0 kcalmol1 is a common method for addressing this problem. The weighting of edge predictions is determined by a measure of uncertainty, such as the standard error of the mean free energy prediction across replicates or an uncertainty estimate derived from bootstrapped subsampling of the simulation data. The estimation of Gbind values is typically accomplished through the use of a weighted-least squares method. The per-ligand Gbound values are calculated by using a reference ligand to determine the value of another of the ligands. The previously indicated data may be used for comparison analysis with experimental binding measurements that have been standardized to the same reference ligand, which makes it possible to evaluate the process of the RBFE.

2. Using Alchemical Free Energy/Machine Learning to Calculate Hydration Free Energies

FEP/ML modelling

The current technique offers a regression model that matches the error an alchemical calculation makes for a given molecule A, defined as:

Goffset(A) = GEXP (A) − GFEP (A), (2.1).

where GFEP (A) is molecule A's al-chemical hydration free energy and GEXP (A) is its experimental one. Five-fold cross-validation across 10 repetitions was used to fit machine-learning models to a training set with prescribed descriptors, resulting in a population Npop of 50 trained models (see methods). All Npop regression models forecast their own G of fset value. Our offset estimator is the arithmetic mean of these offset values, and its accuracy is measured by its standard deviation. Thus, a corrected hydration free energy is:

GFEP/ML(A) = GFEP (A) + G of fset(A) (2.2).

and propagating alchemical and ML term statistical mistakes determines the accuracy of the GFEP/ML(A) estimation.

Data collection

The FreeSolv database, version 0.52, was made available at https://github.com/MobleyLab/FreeSolv. It consists of 642 small, neutral molecules. The GROMACS code used to determine absolute hydration free energy is included in the database.193 In their article, Ramos Matos et al. describe the FEP method. 195 GAFF196 force field, AM1-BCC197 partial charges, and the TIP3P water model were employed in FreeSolv calculations.

All chemicals (n=47) used in the SAMPL4 blinded competition and later added to the FreeSolv database after the aforementioned challenge were removed from the dataset to create the FreeSolvSAMPL4 set.A total of 200 compounds were culled from this group by searching for "SAMPL4 Guthrie" in the database's overview text file's experimental reference column. Mobley 6309289, Mobley 3395921, Mobley 6739648, Mobley 2607611, Mobley 637522, and Mobley 172879 are the six molecules that were purposefully added to the test set by hand. These molecules were included in the SAMPL4 challenge, but the relevant keyword wasn't assigned to them when the FreeSolv database was updated to version 0.52. A total of 595 molecules were used to form the training set. This section will only describe the training set, although all data modification is considered comparable between the training and test sets unless otherwise stated. Python 3.7.4 was used to do the necessary data manipulations.

3. Results & discussion

A. Optimization of protocol on the

training set

The results of the investigation showed that hyperparameters had a major impact on the ML models' validation accuracy. The relatively small size of the training dataset (595 data points) is likely to blame for this phenomena. The research used a hyperparameter optimization approach, which, as shown in Table 2.1, involves adjusting hyperparameters using Bayesian optimization with Gaussian process regression. The relevant approach encompasses noisy and expensive machine learning functions and performs a search inside the hyperparameter space. After 50 iterations, the hyperparameter configuration with the lowest validation error is saved with the trained model. After around 30 iterations, the SVM, RF, and DNN models all showed signs of convergence. It is clear that MLR does not have any tuning-necessary hyperparameters in this situation. As a result, there is no change to the model trained during each SKOPT call, hence there is no change to the validation error either.

All machine learning algorithm hyperparameters are included in Table 1 of this research. SVM, RF, DNN, and MLR are all examples of machine learning models. The configurations of a machine learning model are calculated by multiplying its hyperparameter values.

ML model	Hyperparameter	Range	Tot config n
SVM	C	1e-3, 1e-2, ..., 1e+2	21
	ϵ	1e-3, 1e-2, ..., 1e+2	
	Γ	1e-3, 1e-2, ..., 1e+2	
RF	NumEstimators	1, 2, ..., 1000	9e
	MaxDepth	1, 2, ..., 5	
	MinSamplesSplit	2, 3, ..., 10	
	Bootstrap	True, False	
DNN	ActivationFn	logistic, tanh, relu	3.1e+6
	Solver	lbfgs, sgd, adam	
	Layers*	(100,50),(50,20) , (100,100,50), (100,50,20), (50,20,5)	
	Adam-β1	0.1, 0.2, ..., 0.99	
	Adam-β2	0.1, 0.2, ..., 0.99	
	Adam-ϵ	10e-8, 10e-7, ..., 10e-1	
MLR	No hyperparameters to tune.		1

The findings of the study indicate that the training protocol reveals a comparatively lower degree of fitness of random forests (RF) and multiple linear regressions (MLR) to the training set in comparison to support vector machines (SVM) and deep neural networks (DNN) protocols, as illustrated in Figure 2.2. This outcome is anticipated in the context of Multiple Linear Regression

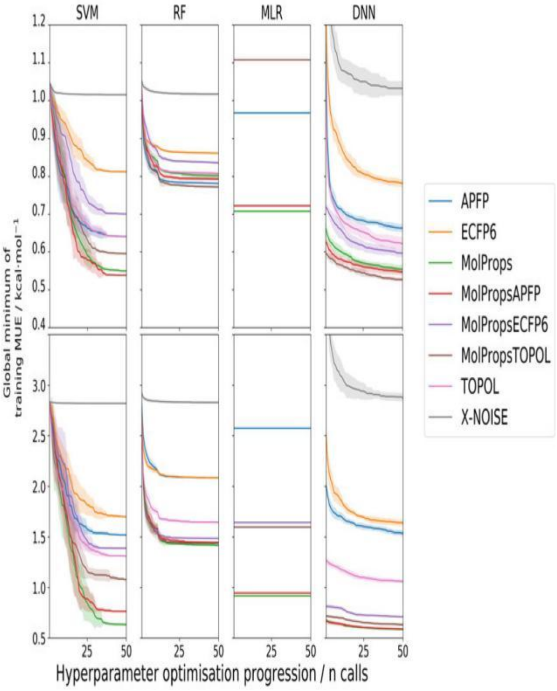

Fig.3: This research optimizes hyperparameters of machine-learning models fitted on two features, namely Goffset (top row) and G (bottom row), for

different kinds of features calculated for compounds in the FreeSolv database. The text calls trained models FEP/ML and pure-ML (ML). The graphic shows the link between hyperparameter calls and training validation mean unsigned error global minima in kcal mol−1. Shaded areas represent the standard deviation of 10 repetitions. Multiple linear regression lines surpass the error range.

(MLR) due to the comparatively uncomplicated nature of the model. Whilst the Random Forest algorithm exhibits greater complexity, it is predominantly tailored towards classification tasks as opposed to regression tasks, owing to its reliance on decision trees. This may account for its tendency to underfit. The present investigation incorporates the algorithm as a means of control.

Various feature sets were employed to determine effective encodings for describing ΔGoffset. A ubiquitous pattern in the performance of feature sets can be discerned among machine learning models. The MolProps and combinatorial feature sets, which include fingerprints appended to MolProps, exhibit a superior fit to the training set compared to standalone fingerprints such as APFP, TOPOL, and ECFP6. Conversely, X-NOISE performs poorly, as anticipated, given that this feature set is derived from random data.

Due to the superior performance of standalone MolProps in comparison to standalone fingerprints, it is probable that the combined feature sets primarily derive their advantages from the more prognostic MolProps component. The empirical evidence indicates that MolProps exhibits superior performance in comparison to other feature sets, implying that certain descriptors (such as molecular weight and polar surface area) incorporated in MolProps exhibit a strong correlation with free energies of hydration. Our empirical findings suggest that the MolProps feature set exhibits superior performance compared to other feature sets in predicting the hydration free energy (ΔG) through the use of pure machine learning models.

4. Conclusions

The present study has exhibited the feasibility of integrating physics-based free energy perturbation (FEP) techniques with data-oriented machine learning approaches for the purpose of forecasting the absolute hydration free energies of minor molecular species. One notable benefit in comparison to FEP is the ability to enhance prediction precision without the need for laborious forcefield parameterization endeavours. The FEP/ML methodology exhibits superior performance in comparison to FEP when considering the training set size, despite both being machine learning techniques. The aforementioned observation holds great importance as it suggests the feasibility of generating predictions for a novel dataset in the absence of any preliminary experimental data, and subsequently transitioning to a Free Energy Perturbation/Machine Learning methodology upon the acquisition of a satisfactory quantity of empirical data points. The aforementioned benefit arises due to the utilization of the FEP/ML methodology, wherein the ML models are solely required to acquire the ability to rectify inaccuracies in the FEP computations. Conversely, in a purely ML-based approach, the models must acquire knowledge pertaining to the physics of hydration. FEP/ML possesses an additional benefit in that it exhibits a comparable level of precision in predicting the hydration free energies of individual compounds as FEP calculations. Conversely, ML-based predictions generated by ensembles of identical models exhibit a more notable degree of variability. Through a retrospective analysis of all submissions made to SAMPL4, it has been observed that the accuracy improvements achieved in Free Energy Perturbation/Machine Learning (FEP/ML) are significant enough to elevate a FEP protocol with a mid-level ranking to a top-ranked submission. Moreover, the enhancements in accuracy are not restricted to a solitary simulation protocol, and several associated Free Energy Perturbation (FEP) methodologies reap the rewards of these rectification factors. This phenomenon is possibly attributable to the observation that various forcefields and software exhibit correlations among their outliers in terms of predicted hydration free energies. 159,216 It is anticipated that the efficacy of the correction terms will diminish as the simulation protocol deviates further from the one employed to produce the training set.

References:

[1] S. Myers and A. Baker, *Nature Biotechnology*, 2001, **19**, 727–730.

[2] L. Nelson, E. Dhimolea and J. M. Reichert, *Nature Reviews Drug Discovery*,2010, **9**, 767–774.

[3] Delany, R. Rappuoli and E. D. Gregorio, *EMBO Molecular Medicine*, 2014,**6**, 708–720.

[4] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham and D. Greyson,*Health Policy*, 2011, **100**, 4–17.

[5] Hughes, S. Rees, S. Kalindjian and K. Philpott, *British Journal of Phar-macology*, 2011, **162**, 1239–1249.

[6] C. H. Wong, K. W. Siah and A. W. Lo, *Biostatistics*, 2018, **20**, 273–286.

[7] Bender and I. Cortês-Ciriano, *Drug Discovery Today*, 2021, **26**, 511–524.

[8] Ha, H. Park, J. Park and S. B. Park, *Cell Chemical Biology*, 2021, **28**,394–423.

[9] S. Fox, S. Farr-Jones, L. Sopchak, A. Boggs, H. W. Nicely, R. Khoury and M. Biros, *SLAS Discovery*, 2006, **11**, 864–869.

[10] Bender and I. Cortes-Ciriano, *Drug Discovery Today*, 2021, **26**, 1040–1052.

[11] Zhao, H. L. Ciallella, L. M. Aleksunes and H. Zhu, *Drug Discovery Today*,2020, **25**, 1624–1638.