Manuscript submitted to **Biophysical***Journal*
Article

Calculation of Protein-Ligand Binding Entropies Using a Rule-Based Molecular Fingerprint

Ali Risheh[1], Alles Rebel[1], Paul S. Nerenberg[2], and Negin Forouzesh[1,*]

[1]Department of Computer Science, California State University, Los Angeles, CA 90032, USA
[2]Kravis Department of Integrated Sciences, Claremont McKenna College, Claremont, CA 91711, USA
[*]Correspondence: neginf@calstatela.edu

ABSTRACT The use of fast *in silico* prediction methods for protein-ligand binding free energies holds significant promise for the initial phases of drug development. Numerous traditional physics-based models (e.g., implicit solvent models), however, tend to either neglect or heavily approximate entropic contributions to binding due to their computational complexity. Consequently, such methods often yield imprecise assessments of binding strength. Machine learning (ML) models provide accurate predictions and can often outperform physics-based models. They, however, are often prone to overfitting, and the interpretation of their results can be difficult. Physics-guided ML models combine the consistency of physics-based models with the accuracy of modern data-driven algorithms. This work integrates physics-based model conformational entropies into a graph convolutional network. We introduce a new neural network architecture (a rule-based graph convolutional network) that generates molecular fingerprints according to predefined rules specifically optimized for binding free energy calculations. Our results on 100 small host-guest systems demonstrate significant improvements in convergence and preventing overfitting. We additionally demonstrate the transferability of our proposed hybrid model by training it on the aforementioned host-guest systems and then testing it on six unrelated protein-ligand systems. Our new model shows little difference in training set accuracy compared to a previous model, but an order of magnitude improvement in test set accuracy. Finally, we show how the results of our hybrid model can be interpreted in a straightforward fashion.

SIGNIFICANCE The significance of this paper lies in the development of a novel hybrid model for predicting protein-ligand binding entropies. Traditional physics-based models often struggle with accurate assessment of entropic contributions due to computational complexity. Machine learning models offer accuracy but can be prone to overfitting and lack interpretability. This work integrates physics-based conformational entropies into an architecture that includes a rule-based graph convolutional layer that generates a new molecular fingerprint suitable for binding calculations. The model demonstrates significant improvements in convergence and preventing overfitting, and it exhibits transferability across different systems. Importantly, this research addresses a critical challenge in drug development, potentially accelerating the initial stages of the process while maintaining accuracy and interpretability in predictions.

1 INTRODUCTION

The drug discovery and development process can span over 12 years and cost over $1 billion (1). The average cost is reported to be $2.6 billion in 2016 (2). The drug search finds and evaluates candidate compounds capable of activating or deactivating specific biological targets through conformational changes. The development process involves delivery, toxicity, testing, etc. (3). High-throughput screening in the early stages of drug discovery uses quick computational methods favoring lower computation time over accuracy (4). Later stages focus on accuracy at the expense of speed.

Binding entropy plays a major role in determining the change in Gibbs free energy (ΔG) of binding. In larger systems, the accurate calculation of entropy becomes critical for determining ΔG due to increased conformational flexibility and solvent interactions (5). In the initial phases of computer-aided screening, implicit solvent free energy calculation methods like Molecular Mechanics/Poisson–Boltzmann Surface Area (MM/PBSA) techniques are frequently employed (6–9). These methods often ignore the contribution of binding entropy due to its computational complexity (5). Note that these implicit solvent models capture solvent entropy contributions in their polar and non-polar terms and are reliant on parameters in their model to estimate this important contribution (10–13).

From an experimental perspective, the determination of binding entropy is challenging, as it requires accurately measuring both the binding free energy and the binding enthalpy (14–16). Consequently, there is a scarcity of datasets containing reliable and accurate experimental values of entropy in the context of binding free energy. For example, PDBBind (17), a reference dataset for binding free energies, does not contain any experimental values for binding entropy. Likewise, binding entropy can be difficult to calculate using computational simulation methods because it requires both complete and accurate sampling of the conformational ensembles of the molecules in their unbound and bound states (14, 15, 18–20). For all of these reasons, it would be valuable to have a fast and accurate computational method for predicting binding entropies that can be trained using relatively sparse experimental data.

Molecular fingerprints are a way of encoding the structure of a molecule for machine learning (ML) methods (21). Due to the varying number of atoms in different molecular systems, it is not feasible to represent atomic information using a vector of fixed size for all molecules. Most ML methods, however, depend on the input data being of a fixed size. Molecular fingerprints provide a solution to this problem by representing the many aspects of a molecular system in a vector of fixed size. Accordingly, fingerprints are context-dependent, as their composition varies depending on their intended usage. For instance, a fingerprint utilized for the prediction of binding free energy might encompass continuous molecular information (e.g., the 3D coordinates of atoms), while a similar fingerprint utilized for predicting toxicity might require only binary features (e.g., whether a given atom is hydrogen). In the nascent stages of molecule encoding, the prevailing methodologies involved the hashing and identification of a distinct identifier for a given molecule based on its constituent atom types, bonds, and overall molecular structure (4, 22). In contrast, more recent solutions utilize differentiable molecular fingerprints generated by ML models (21, 23) that can be used in a wide range of ML algorithms.

Deep learning (DL) models have shown promising results in predicting physical quantities (21, 23). However, one primary concern is the inherent uninterpretability of DL models, commonly referred to as the "black box" problem. Unlike classical physics-based models that can provide explicit equations or rules to explain their predictions, DL models function by utilizing multi-layer weights that are applied to the input data. Consequently, the task of interpreting these layers becomes progressively more difficult when the model architecture comprises more than two layers. Additionally, DL models are prone to errors and can overfit the training data, leading to poor generalization and unreliable predictions when applied to new, unseen examples. Overfitting becomes much more critical for physical quantity prediction problems (e.g., binding entropies) in which the distribution of training data is narrowly clustered about a certain value, resulting in models that return a narrow range of – or even identical – predictions for different input data. These issues necessitate the development of physics-guided ML approaches that can combine the strengths of DL algorithms with the robustness and interpretability of traditional physics-based models. The current state-of-the-art DL models that have shown the highest performance on chemical and physical problems are graph convolutional networks. They are inherently compatible with molecule structures (24), which contain an arbitrary number of bonded atoms. In (21) a graph convolutional network for learning molecular fingerprints was introduced and has been tested on three physical quantity prediction problems: solubility, drug efficacy, and photovoltaic efficiency, all of which are ratio-based quantities. However, its results on regression problems, including entropy calculations, show room for improvement (25).

Physics-guided neural networks (PGNNs) (26–28) provide a systematic framework for combining the scientific knowledge of physics-based models with neural networks to advance scientific discovery. The core idea is to feed the output of a physics-based method into a DL model such that the error of the hybrid model compared to the reference (experiment) is minimized. In this study, we improve the performance of an earlier PGNN model (25) by employing a new molecular fingerprint that is specifically designed and optimized for binding free energy calculations. This paper is organized as follows: we first explain the DL models used in this study, followed by an introduction of the physics-based hybrid model. Next, our novel neural network layer is discussed in detail. The remainder of the paper focuses on the evaluation and interpretation of our model results.

2 MATERIALS AND METHODS

2.1 Deep Learning Models

This section introduces four DL models used in this study (Table 1). In our previous research (25, 27, 28), the goal was to create a hybrid model that combined both atomic features (micro-scale) and other physics-based model outputs (macro-scale) into a single coherent DL architecture. The results showed more robust and accurate models with a high degree of interpretability. The core model that was integrated along with the physics-based model output is a graph convolutional network (GCN) (21). We consider the GCN model as our reference, as it provides a state-of-the-art DL model that is able to transform a molecule into a differentiable fixed-size vector. By feeding this fingerprint to a a simple feed-forward, fully connected network, it is possible to predict physical quantities. All data-driven models take the atom feature matrix (F) and adjacency list as input and predict the entropy y. The physics-guided models, however, consider not only F and the adjacency list, but also incorporate physical parameters (P).

Model Name	Model Description	Featurizer
Graph Convolutional Network (GCN)	This model is an implementation of (21) and is available in DeepChem (29). This model is the reference of this study and our previous studies (25, 27, 28).	Original Graph Convolution Featurizer
Physics-Guided Graph Convolutional Network (PGCN)	This model integrates GCN with physical parameters in the last layer. This model was previously introduced in (27) and has been used in predicting binding free energy	Original Graph Convolution Featurizer
Rule-Based Graph Convolutional Network (RGCN)	The novel model introduced in this paper that utilizes the novel fingerprint to predict the entropy.	Updated Graph Convolution Featurizer
Physics-Guided Rule-Based Graph Convolutional Network (PRGCN)	The novel model introduced in this paper integrated with physical parameters in the last layer.	Updated Graph Convolution Featurizer

Table 1: Descriptions of the different DL models and featurizers used in this study.

2.2 Rule-based Graph Convolutional Layer

In 2015, (21) introduced a new molecular fingerprint created with a GCN. This improvement provided the opportunity to employ DL models in biophysics and biochemistry challenges (30). The GCN model demonstrated accurate predictions in classification problems (31). However, it was limited to applying only one mathematical operation to the features (summation). On the other hand, neither bond types nor interatomic distances were considered in the GCN. These constraints created a barrier for using the model in regression problems. The proposed molecular fingerprint addresses these constraints by using different mathematical operations and considering bond type as a one-hot vector.

A featurizer is used initially to create an atom feature matrix (F) for each molecular system as the input data for the DL models. The original GCN featurizer (32), considers 74 binary features and one continuous. The RGCN featurizer omits 45 atom types to reduce the number of binary features and adds atom mass, atomic number, and 3D coordination of atoms as new features. In this regard, 34 features are considered in the input atom feature matrix (F). As the atom feature matrix (F) is the concatenation of all atom feature vectors (f_i), its size is $N \times M$, where N is the number of atoms and M indicates the number of features we considered for each atom (and consequently the length of the atom feature vector (f)). The core part of the molecular fingerprint is the layer that transforms the input feature matrix (F) into a new continuous matrix. Then, the new matrix is flattened by applying the sigmoid function to each atom feature vector (f') individually and then summing up all the sigmoid output vectors. We call this layer the Rule-Based Graph Convolution Layer (RGCL) because the layer combines the atom features based on different rules: summation over binary features (e.g., atom type) and multiplication for continuous features (e.g., partial charge). As shown in Figure 1, there are two weights defined in this layer. The weight w_{self} is applied on each atom feature vector (f_i) individually. The shape of this weight depends on the output channel size; in this study, the output channel size is 20. Accordingly, the shape of RGCL output matrix will be $N \times 20$, and the shape of w_{self} is 34×20. $w_{neighbor}$ indicates the weight multiplied by the combination of the atom itself and each neighbor feature vector. RGCL transforms the atom feature space into a continuous, differentiable and compact space so the model can train on F'. Next, in the sigmoid transforming layer, each f' is mapped to a vector of length 1024 that is the final fingerprint.

In addition to summation, which was originally used in (21), multiplication and Euclidean distance are used in this research. Specifically, we multiply the partial atomic charges in the RGCL because the strength of the electrostatic interaction between two charged particles depends on the product of their charges (as in Coulomb's Law), rather than the sum. Likewise, the Euclidean distance between two atoms is a more suitable physical feature for the RGCL than a summation of the atoms' coordinates because the sum depends on the origin and orientation of the coordinate axes (i.e., a sum of two coordinates is generally not translationally or rotationally invariant). Table 2 lists all the features and the applied rules.

In (27, 28), we introduced a new DL methodology that considers both atom features (F) and molecular physics parameters (P) together to compensate for physical methods error. The same architecture is applied in this network, and Figure 1 indicates that the entropies calculated by VM2 (see next section) are concatenated to the model variable, which is derived from feeding fingerprint to the stack of dense layers. In this context, the role of the RGCL and fingerprint can be viewed as filling the gap between the experimental data and the (purely) physics-based model.

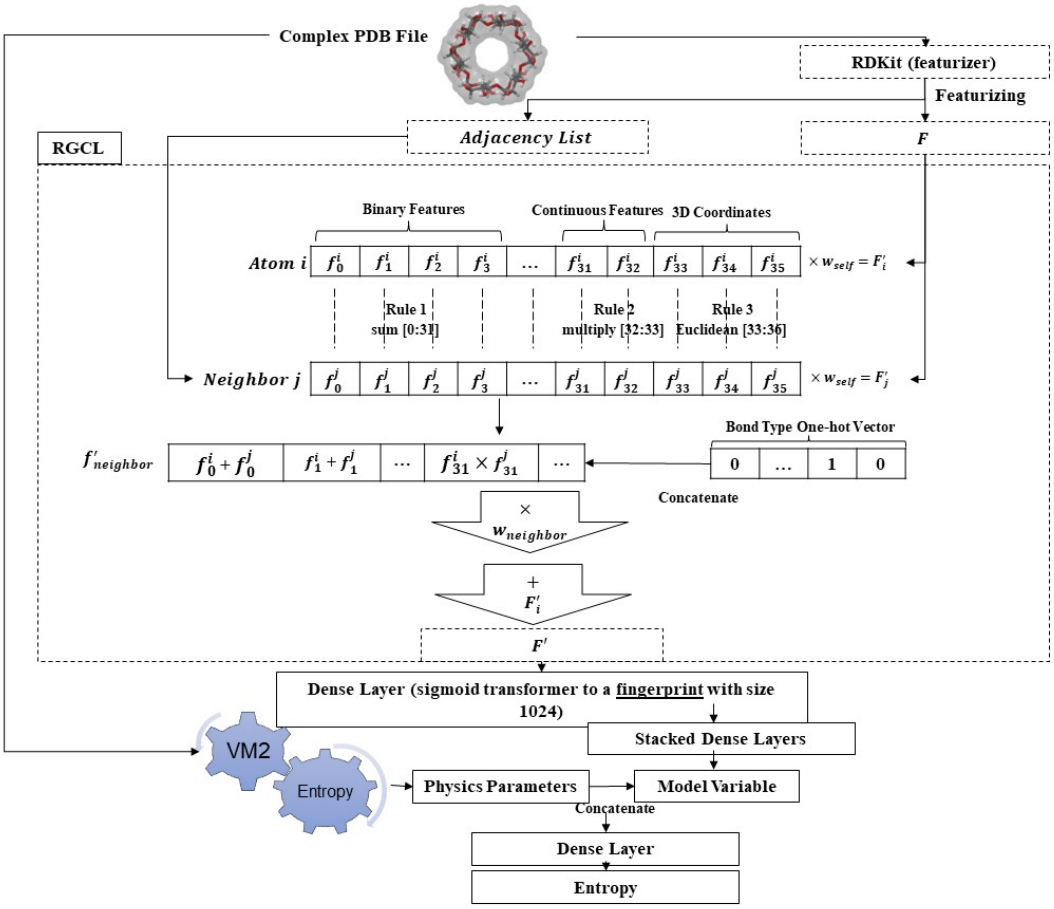

Figure 1: Detailed view of the PRGCN model architecture. The figure depicts the entire physics-guided graph convolutional network designed using the RGCL as the first layer to transform the atom feature matrix (F).

2.3 VeraChem Mining Minima (VM2)

VM2 (33) uses normal mode-based approaches, both to drive conformational searches (mode-distort-minimize) and for the computation of free energies (using a modified harmonic approximation). VM2 utilizes common molecular mechanics (MM) potentials to provide molecular potential energy surfaces. In this work, the General Amber force field (GAFF) (34) was used to parameterize host and guest structures, and the Amber ff99SB (35) force field was used to parameterize BRD4 protein. VM2 uses an implicit solvent model to provide an estimate of the solvation free energy of a given conformation. Following the parameterization, the partial atomic charges are neutralized using VM2's vcharge program. Then, VM2 attempts to find as many conformers as possible, keeping track of calculated energies. This search algorithm (Tork) is designed to be both exhaustive and efficient in exploring the conformational space (36). In host-guest calculations, both the host and guest are fully flexible, but in protein-ligand calculations only part of the protein is flexible (mobile or live atom set), a separate part is present but rigid (fixed or real atom set), and the rest of the protein (beyond a distance cutoff from the ligand) is not included at all. After every search step, a relaxation step is performed to allow the system to move toward the nearest local energy minimum. Once a given iteration of this process has converged to an energy minimum, the resulting conformation is recorded and checked for similarity in energy to others (and combined if similar). VM2 then calculates an approximate local configuration integral that provides that conformation's contribution to the binding free energy. The cumulative binding free energy prediction is updated in an iterative fashion until the conformer search reveals no new energy minima or the search time has been exhausted. The resulting free energy prediction is computationally inexpensive while maintaining reasonable accuracy and correlation with experiment. However, for this work, we focus just on the entropies (i.e., host, guest, and host-guest complex) predicted by VM2. It should be noted that the entropies reported by VM2 are conformational entropies of the solute molecules only and do not include any solvent entropies due to the use of implicit solvation.

Feature name	Abbreviation	Type	Operand (rule)
Atom type C	C	Binary	sum
Atom type N	N	Binary	sum
Atom type O	O	Binary	sum
Atom type S	S	Binary	sum
Atom type F	F	Binary	sum
Number of bonded Hydrogen	H	Integer	sum
Atom type unknown	Unk	Binary	sum
Number of atoms bonded	Atom degree	One-hot vector [0-3]	sum
Implicit valence	Implicit Valence	One-hot vector [0-5]	sum
Number of radical electrons	Electrons	Integer	sum
Atom formal charge	Formal charge	Integer	multiplication
Atom hybridization (sp, sp^2, ...)	SP	One-hot vector [0-4]	sum
Is atom aromatic	Aromatic	Binary	sum
Atom mass	Mass	Float	sum
Atomic number	Atomic number	Integer	sum
Atomic 3D position	position	3 Integers	Euclidean distance
Bond type (single, double, aromatic, other)	Bond	One-hot vector [0-3]	N/A*

Table 2: Micro-scale features considered for each atom in the featurizer. Where applicable, the number of one-hot vector elements is indicated to the right. The bond type feature is not considered an individual atomic feature, as it represents a shared characteristic between two bonded atoms. Instead, it is concatenated with combined features and subsequently multiplied by $W_{neighbor}$.

2.4 Datasets

We collected experimental data on 100 host-guest systems from a variety of sources, including the SAMPL challenge exercises, which are held periodically to assess community progress in the *in silico* prediction of binding affinities and related quantities (37–47). The host molecules found in these studies are octa-acid (OA), cucurbituril (CBn), or cyclodextrin molecules (Figure 2), and the guest molecules are typically small organic molecules. Additionally, we used data obtained for the BRD4 protein complexed with six different ligands (34, 48–59) to test the transferability of our models to larger structures. Binding entropies are calculated from the experimental data using the equation $-T\Delta S = \Delta G - \Delta H$, as entropy is not measured directly in the experimental studies. The probability density of the entropy values and the corresponding kernel density estimation are presented in Figure 3. The host-guest dataset is divided at the rate of 80:20 for training and testing the DL models, respectively.

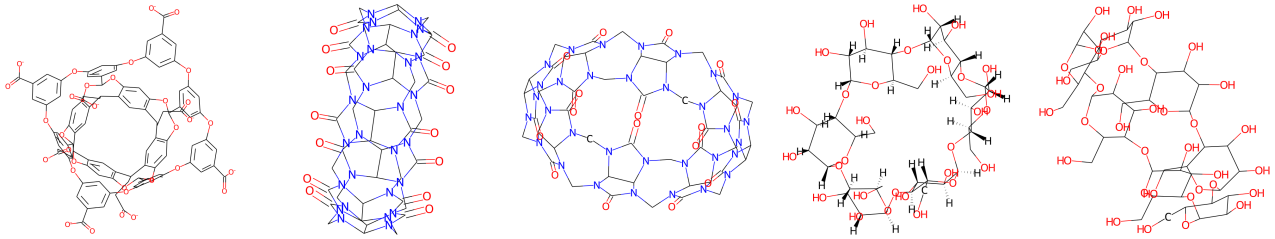

Figure 2: SAMPL hosts used in this research (from left to right): OA, CB7, CB8, α-cyclodextrin, β-cyclodextrin.

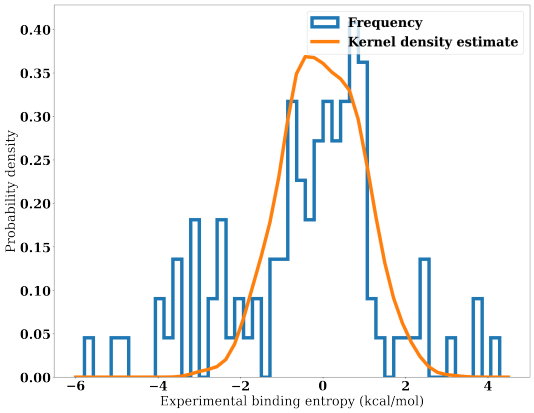

Figure 3: Distribution of the experimental binding entropies from the host-guest dataset.

3 RESULTS AND DISCUSSION

3.1 Model Training and Evaluation

We first compared the performance of the four models using a 5-fold cross-validation over the host-guest systems. Figure 4 depicts the convergence steps in the training process through monitoring of the training loss (RMSE). It is immediately apparent that the data-driven models (GCN and RGCN) have smaller RMSEs than the physics-guided models (PGCN and PRGCN). Moreover, the data-driven models demonstrate only small improvements in RMSE – suggestive of overfitting – while the physics-guided models start with RMSEs that are comparable to the predictions made by the physics-based VM2 model (13.83 kcal/mol RMSE over the entire host-guest dataset) and then subsequently converge to RMSEs that are comparable, albeit slightly higher than those of the data-driven models. This suggests that the "guidance" from a physics-based model is able to effectively regularize the DL models and reduce the potential for overfitting of the training data.

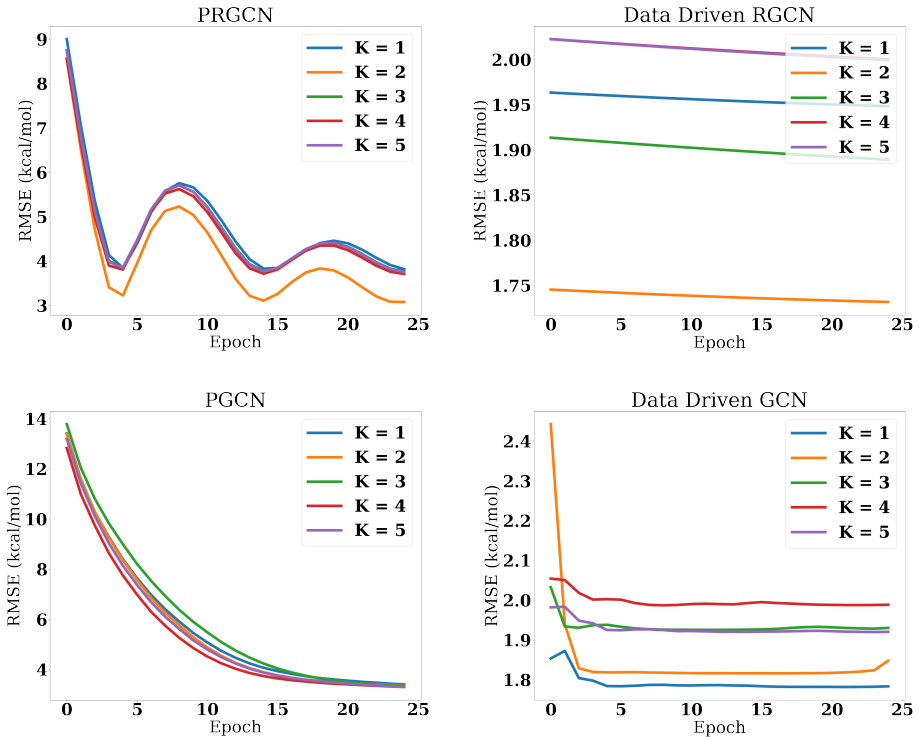

Figure 4: Training losses for each of the tested models on a 5-fold cross-validation set of the host-guest binding data.

In order to gain a deeper understanding of the obstacles inherent in physical quantity prediction, it is instructive to examine the predictive performance of these models on both the training and test sets (following the completion of the training process). This post-training analysis can provide valuable insights into the efficacy and robustness of a given model. Figure 5 shows that the data-driven models – regardless of fingerprint used – yield predictions that vary little, if at all, with different input data. The likely reason for this behavior is that the models' goal is to minimize the loss function, and an optimal RMSE can be obtained by having the predictions close to the mean of the training data. Indeed, the predicted outputs of these models appear to be quite close to the mean binding entropy of the full host-guest dataset (see the kernel density estimate in Figure 3). This provides further evidence that these models are overfitting the training data. In contrast, the physics-guided models yield predictions that are of approximately the same range as the experimental data, albeit with considerable scatter.

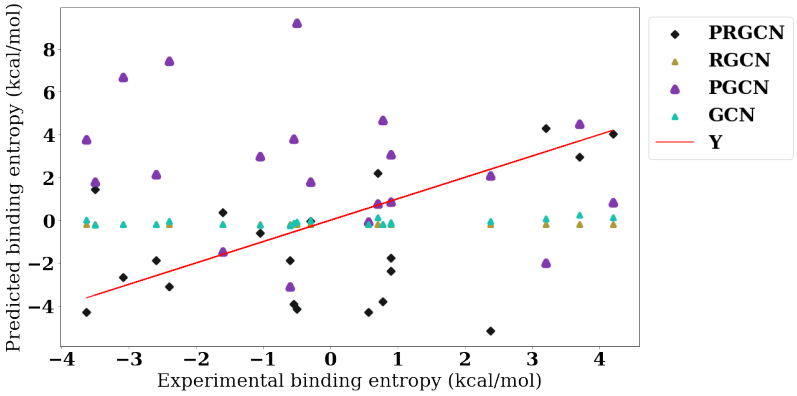

Figure 5: Comparison of experimental and predicted binding entropies for the $K = 1$ test set. The red line (labeled Y) represents perfect agreement with the experimental data.

The above observations are corroborated by examining the test set RMSEs and combined training and test set correlation coefficients shown in Table 3. Indeed, we find that the RGCN model predicts the *exact* same binding entropy, regardless of the input data, as both correlation coefficients are zero. The GCN model fares better, with the largest positive correlation coefficient on average, but we again note that its predictions vary by only very small amounts from the training set means (Figure 5). Both physics-guided models have correlation coefficients that are mildly positive (Table 3). Moreover, we find that the PRGCN has a more consistent RMSE across the training and test sets compared to the PGCN, as well as a somewhat larger Pearson correlation coefficient (Table 3). In contrast, VM2, the physics-based model that provides input for both the PRGCN and PGCN models, yields binding entropy predictions that have an RMSE of 13.83 kcal/mol across the entire dataset, with a Pearson correlation of -0.06 and a Kendall rank correlation of -0.06.

Model	Training loss	Test loss	Pearson correlation	Kendall rank correlation
PRGCN (physics+data)	3.80 ± 0.27	3.98 ± 0.93	0.18 ± 0.07	0.10 ± 0.07
RGCN (data-driven)	1.91 ± 0.09	1.88 ± 0.36	0.00	0.00
PGCN (physics+data)	3.14 ± 0.03	4.88 ± 0.12	0.11 ± 0.07	0.08 ± 0.08
GCN (data-driven)	1.89 ± 0.07	1.91 ± 0.35	0.33 ± 0.19	0.20 ± 0.07

Table 3: Training and test results from 5-fold cross-validation on 100 host-guest systems. The training and test losses are given in kcal/mol. The correlation coefficients are computed for the combined training and test set of each fold and then averaged.

These results highlight the potential pitfall of overfitting the training data when using DL models in physical quantity regression. There are many solutions to overcome the problem of overfitting (60, 61), but the solution we have presented here is to feed predictions from a computationally inexpensive physics-based model to the DL model.

3.2 Model Transferability

We next assessed the transferability of the physics-guided model using the new fingerprint (PRGCN) against that of the physics-guided model using the previous fingerprint (PGCN). To do this, we trained it on all 100 host-guest structures and then tested it on six structures of a large protein (BRD4) bound to different ligands. From these results (Table 4), we conclude that the PRGCN model is more robust and can successfully learn how to predict physical quantities even with small and few structures. The rationale behind this relies on two fundamental differences between the RGCN and the GCN. First, the RGCN has many fewer (order of 1/100) weights compared to the GCN, and this should help prevent overfitting of the model. Second, the RGCN utilizes different mathematical operations applied to the features, in contrast to the GCN, which only applies summation over features. Therefore, the RGCN may be more able to find physical equations/relationships within the network weights. As mentioned above, the RGCN is primarily designed for regression, and our goal is to make a model capable of learning physical calculations in depth so it will be robust across different molecular systems.

Model	Training loss	Test loss
PRGCN	2.77	2.95
PGCN	2.12	24.66

Table 4: Results of training the physics-guided models on the full 100 host-guest systems and then testing on 6 protein-ligand systems. The training and test losses are given in kcal/mol.

3.3 Model Interpretation

DL models are often referred to as "black boxes" because interpreting what the hidden layers are responding to within the input data and how they are contributing to a given model's predictions can be challenging. The ultimate goal of research such as the current work is to use DL to discover promising candidate drug molecules for human health. Because an incorrect prediction in this field would be costly – in terms of additional time spent on research, financial resources spent on development and testing, and the like – there is considerable value in having some sense of how a given model arrives at its predictions. Here we explore different ways of interpreting how our physics-guided models make their predictions, with a particular focus on the more transferable PRGCN model.

As mentioned before, the core part of our model predictions relies on the graph convolution layer that transforms the input feature matrix (F) to a new space (F'). Therefore, we focus specifically on the GCN and RGCN layers. The GCN layer (GCL) proposed by (21) defines a weight per neighbor. Therefore, in our study when considering 10 nearby atoms and 64 as the output channel, there are $10 \times 75 \times 64$ weights defined in the GCL. In contrast, the RGCN layer (RGCL) assumes only one weight matrix for an atom itself and one weight matrix for any bonded atoms. In this study, the output channel was 20, which results in $2 \times 38 \times 20$ weights defined in the RGCL. Overall, there are 102144 weights in the GCL and just 1440 weights in the RGCL. The comparison between number of weights in both layers outlines the importance of layer architecture. Although the RGCL contains fewer weights, we find that it is more robust and accurate . This is in large part why we developed a new neural network architecture instead of using an off-the-shelf architecture.

One basic interpretation of a neural network model can come from the magnitudes of the weights. Each atom feature is multiplied by multiple weights, and the magnitudes of the weights implies the importance of the feature from the model's point of view. It is notable that this assumption is applicable in this study's layers because they transform the feature matrix (F) using the multiplication of weights to features. The GCL (21) assumes a specific weight for different number of neighbors. In this regard, it is not straightforward to extract the magnitude of overall weights applied to each feature, and therefore we only considered the weight that is applied to the atom itself. In contrast, the RGCL has only two weights that are applied to the atom itself and its neighbors' features, respectively. This simplification not only improved the result considerably but also enhanced the layer interpretation. Figure 6 shows the magnitude of each feature in each layer. Although both models have slightly different trends among their feature weights, the variation of weights in the PRGCN model is substantially greater than in the PGCN model. The precise meaning of these numbers is not explicit, but it suggests that the RGCL may be better able to distinguish between different features than the GCL.

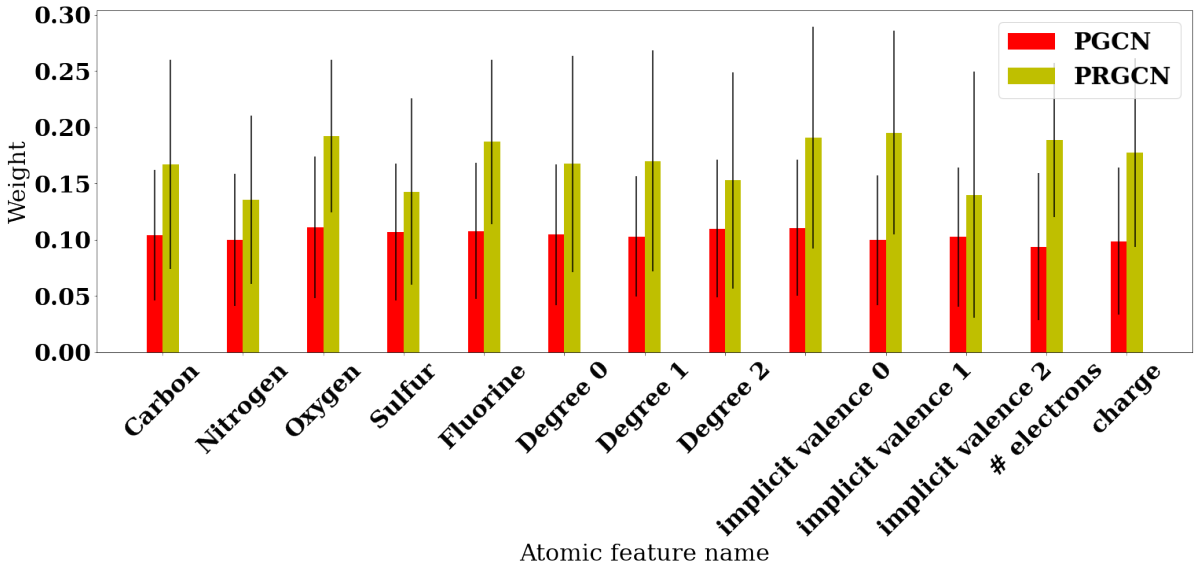

Figure 6: Mean and standard deviation of weights applied to atom features individually. The numbers in chart show the absolute mean value of weights applied to each feature.

To further expand our understanding of the model, we took an altogether different approach. We repeatedly trained the PRGCN model for 25 epochs, but excluded a different input feature each time. Figure 7 demonstrates that all of the features are important to the PRGCN's model accuracy, as removing any of them yields an increase of RMSE (test loss) from 4.04 to ~4.15 kcal/mol. Moreover, the differences in RMSE between removing any one of the features arfe relatively small (~0.001 kcal/mol) compared to the overall increase in RMSE (~0.1 kcal/mol). We were unable to discern any strong correlation between the results shown in Figures 6 and 7, however, as there are approximately as many instances of a larger feature weight corresponding to a larger increase in RMSE when that feature is removed as there are of the converse.

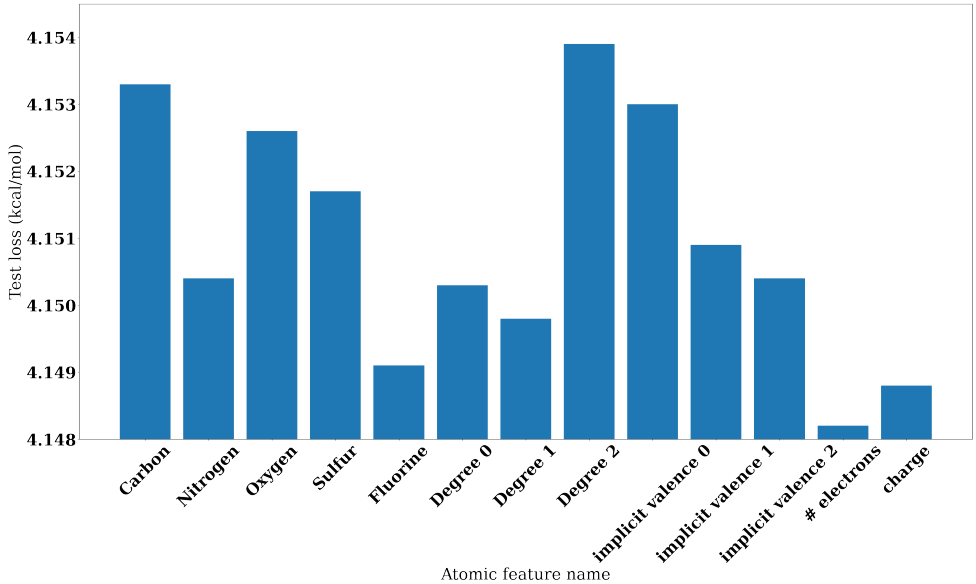

Figure 7: Test set loss (in kcal/mol) of the PRGCN model after excluding each feature during training. Before removing any features, the baseline test set loss is 4.04 kcal/mol.

4 CONCLUSION

Purely data-driven models can outperform physics-based models but are always at risk of overfitting the training data. Another critique is that the results of data-driven models often lack meaningful interpretation due to the "black box" characteristics of ML. When data-driven models are coupled with physics-based models, the resulting hybrid model inherits high accuracy from the former and interpretability from the latter. Earlier works examined the application of such hybrid models to predict binding enthalpies, the other primary component of binding free energies. This work differs by looking at the application of hybrid models to binding entropy predictions, using a new molecular fingerprint and computationally inexpensive physics-based entropy predictions. Our study suggests that our novel neural network layer could overcome the challenges concerning regression problems in biochemistry and biophysics. The results of this model show two main areas of improvement relative to previous models. First, the model's interpretability is improved due to a smaller number of weights and also integration with physics-based model output. Second, based on the results from testing the model on large molecular systems never used in training, the model's transferability is also improved.

There are many possible avenues for future development of this model. For example, the physics parameters that are concatenated to the model variable could be optimized. One possible approach would be to add different physics parameters and check the accuracy and parameter coefficients (model weights multiplied on the parameters) to see if they are used efficiently and correctly in the prediction. In addition, the loss function plays a critical role in determining the weights of the model. In this regard, providing more complex loss functions that consider physical consistency of the model could improve the model's results and interpretability significantly.

In conclusion, DL models hold promise for speeding up the drug discovery process by refining the accuracy of physics-based models, especially those that are faster but less accurate than *ab initio* methods. Indeed, the network architecture introduced here could be utilized on a large number of structures and predict different components of binding free energy or predict binding free energy directly. Nonetheless, for these DL models to be considered a viable approach in this field, it is crucial to improve both their accuracy and their ability to work seamlessly with existing methods with a high degree of interpretability.

Data Availability Statement: Both the data and model information can be found publicly available online at dataset in GitHub and models in GitHub.

AUTHOR CONTRIBUTIONS

N.F. and A.R. (Author1) designed the research. A.R. (Author1) and A.R. (Author2) carried out simulations and analyzed the data. A.R. (Author1), P.S.N., and N.F. wrote the article. P.S.N. and N.F. provided resources and supervision.

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health under award number R16GM146633 (to N.F.) and by the National Science Foundation under award number 2216858 (to N.F. and P.S.N.). We thank the Extreme Science and Engineering Discovery Environment (XSEDE) for providing computational resources used in this research. We appreciate VeraChem LLC for providing us access to VM2 entire software suite to perform experiments, validation of published works, and technical support. We acknowledge Ali Falahati for his assistance in running simulations related to the model interpretation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. DiMasi J.A., S. A., Feldman L., and W. A., 2010. Trends in risks associated with new drug development: success rates for investigational drugs. *Clin Pharmacol Ther* 87:272–277.

2. DiMasi J.A., G. H., and H. R.W, 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Health Econ* 47:20–33.

3. W.L.Jorgensen, 2004. The many roles of computation in drug discovery. *Science* 303:1813–1818.

4. Rogers, D., and M. Hahn, 2010. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* 50:742–754. https://doi.org/10.1021/ci100050t, pMID: 20426451.

5. Goethe, M., J. Gleixner, I. Fita, and J. M. Rubi, 2018. Prediction of Protein Configurational Entropy (Popcoen). *Journal of Chemical Theory and Computation* 14:1811–1819. https://doi.org/10.1021/acs.jctc.7b01079, pMID: 29351717.

6. Forouzesh, N., and N. Mishra, 2021. An effective MM/GBSA protocol for absolute binding free energy calculations: a case study on SARS-CoV-2 spike protein and the human ACE2 receptor. *Molecules* 26:2383.

7. Mishra, N., and N. Forouzesh, 2022. Protein-Ligand Binding with Applications in Molecular Docking. *In* Algorithms and Methods in Structural Bioinformatics, Springer, 1–16.

8. Forouzesh, N., 2020. Binding Free Energy of the Novel Coronavirus Spike Protein and the Human ACE2 Receptor: An MMGB/SA Computational Study. *In* Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 1–7.

9. Panday, S. K., and E. Alexov, 2022. Protein–protein binding free energy predictions with the MM/PBSA approach complemented with the gaussian-based method for entropy estimation. *ACS omega* 7:11057–11067.

10. Decherchi, S., M. Masetti, I. Vyalov, and W. Rocchia, 2015. Implicit solvent methods for free energy estimation. *European Journal of Medicinal Chemistry* 91:27–42. https://www.sciencedirect.com/science/article/pii/S0223523414007995, molecular Dynamics: New Advances in Drug Discovery.

11. Forouzesh, N., S. Izadi, and A. V. Onufriev, 2017. Grid-based surface generalized Born model for calculation of electrostatic binding free energies. *Journal of chemical information and modeling* 57:2505–2513.

12. Onufriev, A., 2010. Continuum electrostatics solvent modeling with the generalized Born model. *Modeling Solvent Environments: Applications to Simulations of Biomolecules* 127–165.

13. Onufriev, A. V., and S. Izadi, 2018. Water models for biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 8:e1347.

14. Linkuvienė, V., G. Krainer, W.-Y. Chen, and D. Matulis, 2016. Isothermal titration calorimetry for drug design: Precision of the enthalpy and binding constant measurements and comparison of the instruments. *Analytical Biochemistry* 515:61–64. https://www.sciencedirect.com/science/article/pii/S000326971630327X.

15. Jarmoskaite, I., I. AlSadhan, P. P. Vaidyanathan, and D. Herschlag, 2020. How to measure and evaluate binding affinities. *eLife* 9:e57264. https://doi.org/10.7554/eLife.57264.

16. Gohlke, H., and G. Klebe, 2002. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie International Edition* 41:2644–2676.

17. Wang, R., X. Fang, Y. Lu, and S. Wang, 2004. The PDBbind Database: Collection of Binding Affinities for ProteinLigand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry* 47:2977–2980. https://doi.org/10.1021/jm030580l.

18. Homeyer, N., F. Stoll, A. Hillisch, and H. Gohlke, 2014. Binding free energy calculations for lead optimization: assessment of their accuracy in an industrial drug design context. *Journal of chemical theory and computation* 10:3331–3344.

19. Chang, C.-e. A., W. Chen, and M. K. Gilson, 2007. Ligand configurational entropy and protein binding. *Proceedings of the National Academy of Sciences* 104:1534–1539.

20. Ben-Shalom, I. Y., S. Pfeiffer-Marek, K.-H. Baringhaus, and H. Gohlke, 2017. Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations. *Journal of chemical information and modeling* 57:170–189.

21. Duvenaud, D., D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints.

22. Morgan, H. L., 1965. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* 5:107–113. https://doi.org/10.1021/c160017a018.

23. Kearnes, S., K. McCloskey, M. Berndl, V. Pande, and P. Riley, 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* 30:595–608. https://doi.org/10.1007%2Fs10822-016-9938-8.

24. Zhang, S., H. Tong, J. Xu, and R. Maciejewski, 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks* 6:11. https://doi.org/10.1186/s40649-019-0069-y.

25. Rebel, A., A. Risheh, N. Massoudian, and N. Forouzesh, 2022. Calculating the Binding Entropy of Host-Guest Systems with Physics-Guided Neural Networks. *In* 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 3478–3485.

26. Daw, A., A. Karpatne, W. Watkins, J. Read, and V. Kumar, 2017. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. https://arxiv.org/abs/1710.11431.

27. Cain, S., A. Risheh, and N. Forouzesh, 2021. Calculation of Protein-Ligand Binding Free Energy Using a Physics-Guided Neural Network. *In* 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2487–2493.

28. Cain, S., A. Risheh, and N. Forouzesh, 2022. A Physics-Guided Neural Network for Predicting Protein-Ligand Binding Free Energy: From Host Guest Systems to the PDBbind Database. *Biomolecules* 12. https://www.mdpi.com/2218-273X/12/7/919.

29. 2016. Democratizing Deep-Learning for Drug Discovery, Quantum Chemistry, Materials Science and Biology. https://github.com/deepchem/deepchem.

30. Coley, C. W., R. Barzilay, W. H. Green, T. S. Jaakkola, and K. F. Jensen, 2017. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *Journal of Chemical Information and Modeling* 57:1757–1772. https://doi.org/10.1021/acs.jcim.6b00601.

31. Chowdhury, A., S. Srinivasan, A. Mukherjee, S. Bhowmick, and K. Ghosh, 2023. Improving Node Classification Accuracy of GNN through Input and Output Intervention. *ACM Trans. Knowl. Discov. Data* 18. https://doi.org/10.1145/3610535.

32. 2016. Duvenaud graph convolutions construct a vector of descriptors for each atom. https://deepchem.readthedocs.io/en/latest/api_reference/featurizers.html#convmolfeaturizer.

33. VeraChem, 2019. Free Energy Calculations by Mining Minima¶. https://www.verachem.com/vm2_free_energy_calculation_by_mining_minima/.

34. Wang, J., R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, 2004. Development and testing of a general amber force field. *Journal of Computational Chemistry* 25:1157–1174. https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20035.

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

36. Chang, C.-E., and M. K. Gilson, 2003. Tork: Conformational Analysis Method for Molecules and Complexes. *Journal of Computational Chemistry* 24:1987–1998.

37. Muddana, H. S., C. D. Varnado, C. W. Bielawski, A. R. Urbach, L. Isaacs, M. T. Geballe, and M. K. Gilson, 2012. Blind prediction of host-guest binding affinities: a new SAMPL3 challenge. *J. Comput. Aided Mol. Des.* 26:475–487.

38. Muddana, H. S., A. T. Fenley, D. L. Mobley, and M. K. Gilson, 2014. The SAMPL4 host-guest blind prediction challenge: an overview. *J. Comput. Aided Mol. Des.* 28:305–317.

39. Yin, J., N. M. Henriksen, D. R. Slochower, M. R. Shirts, M. W. Chiu, D. L. Mobley, and M. K. Gilson, 2017. Overview of the SAMPL5 host–guest challenge: Are we doing better? *Journal of Computer-Aided Molecular Design* 31:1–19.

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

41. Mobley, D. L., G. Heinzelmann, N. M. Henriksen, and M. K. Gilson, 2017. Predicting binding free energies: Frontiers and benchmarks (a perpetual review).

42. MobleyLab. MobleyLab/benchmarksets: Benchmark sets for binding free energy calculations: Perpetual Review Paper, discussion, datasets, and standards. https://github.com/MobleyLab/benchmarksets.

43. Wickstrom, L., P. He, E. Gallicchio, and R. M. Levy, 2013. Large Scale Affinity Calculations of Cyclodextrin Host–Guest Complexes: Understanding the Role of Reorganization in the Molecular Recognition Process. *Journal of Chemical Theory and Computation* 9:3136–3150. https://doi.org/10.1021/ct400003r.

44. Zhang, H., C. Yin, H. Yan, and D. van der Spoel, 2016. Evaluation of Generalized Born Models for Large Scale Affinity Prediction of Cyclodextrin Host–Guest Complexes. *Journal of Chemical Information and Modeling* 56:2080–2092. https://doi.org/10.1021/acs.jcim.6b00418.

45. Yin, J., N. M. Henriksen, D. R. Slochower, and M. K. Gilson, 2017. The SAMPL5 host–guest challenge: computing binding free energies and enthalpies from explicit solvent simulations by the attach-pull-release (APR) method. *Journal of Computer-Aided Molecular Design* 31:133–145. https://doi.org/10.1007/s10822-016-9970-8.

46. Bosisio, S., A. S. J. S. Mey, and J. Michel, 2017. Blinded predictions of host-guest standard free energies of binding in the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design* 31:61–70. https://doi.org/10.1007/s10822-016-9933-0.

47. Rekharsky, M. V., M. P. Mayhew, R. N. Goldberg, P. D. Ross, Y. Yamashoji, and Y. Inoue, 1997. Thermodynamic and Nuclear Magnetic Resonance Study of the Reactions of α- and β-Cyclodextrin with Acids, Aliphatic Amines, and Cyclic Alcohols. *The Journal of Physical Chemistry B* 101:87–100. https://doi.org/10.1021/jp962715n.

48. Lucas, X., D. Wohlwend, M. gle, K. Schmidtkunz, S. Gerhardt, R. le, M. Jung, O. Einsle, and S. nther, 2013. 4-Acyl pyrroles: mimicking acetylated lysines in histone code reading. *Angew Chem Int Ed Engl* 52:14055–14059.

49. O'Boyle, N. M., M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, 2011. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* 3:33. https://doi.org/10.1186/1758-2946-3-33.

50. Jakalian, A., D. B. Jack, and C. I. Bayly, 2002. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* 23:1623–1641. https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.10128.

51. Trott, O., and A. J. Olson, 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31:455–461.

52. Maier, J. A., C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* 11:3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

53. Joung, I. S., and T. E. Cheatham III, 2008. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *The Journal of Physical Chemistry B* 112:9020–9041. https://doi.org/10.1021/jp8001614.

54. Vidler, L. R., P. Filippakopoulos, O. Fedorov, S. Picaud, S. Martin, M. Tomsett, H. Woodward, N. Brown, S. Knapp, and S. Hoelder, 2013. Discovery of Novel Small-Molecule Inhibitors of BRD4 Using Structure-Based Virtual Screening. *Journal of Medicinal Chemistry* 56:8073–8088. https://doi.org/10.1021/jm4011302.

55. Fish, P. V., P. Filippakopoulos, G. Bish, P. E. Brennan, M. E. Bunnage, A. S. Cook, O. Federov, B. S. Gerstenberger, H. Jones, S. Knapp, B. Marsden, K. Nocka, D. R. Owen, M. Philpott, S. Picaud, M. J. Primiano, M. J. Ralph, N. Sciammetta, and J. D. Trzupek, 2012. Identification of a Chemical Probe for Bromo and Extra C-Terminal Bromodomain Inhibition through Optimization of a Fragment-Derived Hit. *Journal of Medicinal Chemistry* 55:9831–9837. https://doi.org/10.1021/jm3010515.

56. Filippakopoulos, P., S. Picaud, O. Fedorov, M. Keller, M. Wrobel, O. Morgenstern, F. Bracher, and S. Knapp, 2012. Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family. *Bioorg Med Chem* 20:1878–1886.

57. Picaud, S., C. Wells, I. Felletar, D. Brotherton, S. Martin, P. Savitsky, B. Diez-Dacal, M. Philpott, C. Bountra, H. Lingard, O. Fedorov, S. ller, P. E. Brennan, S. Knapp, and P. Filippakopoulos, 2013. RVX-208, an inhibitor of BET transcriptional regulators with selectivity for the second bromodomain. *Proc Natl Acad Sci U S A* 110:19754–19759.

58. Filippakopoulos, P., J. Qi, S. Picaud, Y. Shen, W. B. Smith, O. Fedorov, E. M. Morse, T. Keates, T. T. Hickman, I. Felletar, M. Philpott, S. Munro, M. R. McKeown, Y. Wang, A. L. Christie, N. West, M. J. Cameron, B. Schwartz, T. D. Heightman, N. La Thangue, C. A. French, O. Wiest, A. L. Kung, S. Knapp, and J. E. Bradner, 2010. Selective inhibition of BET bromodomains. *Nature* 468:1067–1073.

59. Gehling, V. S., M. C. Hewitt, R. G. Vaswani, Y. Leblanc, A. Côté, C. G. Nasveschuk, A. M. Taylor, J.-C. Harmange, J. E. Audia, E. Pardo, S. Joshi, P. Sandy, J. A. Mertz, R. J. Sims III, L. Bergeron, B. M. Bryant, S. Bellon, F. Poy, H. Jayaram, R. Sankaranarayanan, S. Yellapantula, N. Bangalore Srinivasamurthy, S. Birudukota, and B. K. Albrecht, 2013. Discovery, Design, and Optimization of Isoxazole Azepine BET Inhibitors. *ACS Medicinal Chemistry Letters* 4:835–840. https://doi.org/10.1021/ml4001485.

60. Ying, X., 2019. An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series* 1168:022022. https://dx.doi.org/10.1088/1742-6596/1168/2/022022.

61. Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15:1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.