# Predicting Macrocyclic Molecular Recognition with Machine Learning

Anthony Tabet,[†,‡,¶,§] Thomas Gebhart,[‖] Guanglu Wu,[¶] Charlie Readman,[¶]
Merrick Pierson Smela,[¶] Vijay K. Rana,[¶] Cole Baker,[⊥] Harry Bulstrode,[§]
Polina Anikeeva,[‡] David H. Rowitch,[§] and Oren A. Scherman[*,¶]

†*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA USA*
‡*Department of Materials Science & Engineering, Massachusetts Institute of Technology, Cambridge, MA USA*
¶*Melville Laboratory for Polymer Synthesis, Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK*
§*Department of Paediatrics, Addenbrooke's Hospital, University of Cambridge, Hills Road, Cambridge CB2 0QQ, UK*
‖*Department of Computer Science, University of Minnesota, Minneapolis, MN USA*
⊥*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA USA*

Received August 12, 2019; E-mail: oas23@cam.ac.uk

***Abstract:*** DFT calculations are used as training data to predict equilibrium binding constants of small molecules to cucurbit[7]uril (CB[7]) with kernel-based support vector machine learning. This trained algorithm was then used to predict the binding of two promising small molecule drugs in the clinic against pediatric low grade glioma, TAK-580 and Selumetinib. The algorithm predicted strong binding for TAK-580 and poor binding for Selumetinib. These results were experimentally validated. It was also discovered that the slightly larger homologue cucurbit[8]uril (CB[8]) is partial to Selumetinib, suggesting an opportunity for tunable release kinetics by introducing different concentrations of CB[7] or CB[8] into a system such as a hydrogel depot for local drug delivery. We also qualitatively demonstrated that these two drugs have different therapeutic windows and may have utility in combination against low grade gliomas. Finally, mass transfer simulations were performed to show how CB[7] can independently tune the release of TAK-580 across time scales from seconds to a year without changing the kinetics of Selumetinib. This work shows how machine learning may prove valuable in the development of drug-delivery systems for combination therapies and the field of supramolecular chemistry more broadly.

The application of machine learning in biology and chemistry has received heightened attention in recent years.[1–5] This rapidly expanding paradigm is exciting due to the potential of data science to improve small molecule drug discovery, identify more efficient synthetic pathways, create proteins with greater binding affinity to specific substrates, and other applications. One such application that has not yet been explored is predicting the molecular recognition of small molecules with macrocycles.

Cucurbiturils are a class of symmetric macrocycles that have applications within drug delivery, biosensing, catalysis, and energy.[6,7] These macrocycles have many advantages over their non-symmetric counterparts such as cyclodextrins, including temperature stability and robustness at acidic and basic pH values,[6] such as those that occur naturally in physiology. The use of cucurbiturils to change the release kinetics or pharmacokinetics of drugs has been previously reported for chemotherapies such as temozolomide.[8,9] Cucurbituril acts as a competitive substrate for the active ingredient; such a phenomena can reduce the effective concentration and increase the half life of biologic and hydrophobic small molecule drugs.[10] Predicting whether a molecule will bind to any cucurbituril, in particular cucurbit[7]uril, *a priori* could be an invaluable tool in developing new chemical or material systems.[11]

In this work, we report the prediction of 1:1 complexation of small organic molecules with cucurbit[7]uril. Finding no comprehensive, compiled body of data that could be used for regression, we first created one. We also report the utility of this regression in predicting the binding of two new small molecule drugs that have received promising results in the clinic and verify these predictions with experimental data. Finally, we provide a qualitative example of the potential use of these predictions in developing cocktail drug therapies against a pediatric low grade glioma cell model.

The principle challenge for any machine learning application is in building a sufficiently large training data set that approximates the entire problem domain with as little bias as possible.[12] We performed density functional theory (DFT) simulations on 146 unique molecules and 196 total different solution conditions such as variable ionic strength (Table S1).[13,14] Of these, 145 were good guests for CB[7] (obtained from the literature[6]). Seeing a lack of negative controls, we also synthesized and/or tested three molecules that could not bind to CB[7] and set these undetectable binding events to output values of 0 to not skew the algorithm with extreme values (Fig. S1-11).

Critical to the binding affinity of molecules with CB[7] are the size, aromaticity, and charge of the guest. Other, non-intrinsic parameters such as solution temperature, pH, salt and/or buffer concentration may also effect the equilibrium binding constant.[6,15] We sought to capture both intrinsic and environmental properties of the binding event as potential predictive features (Fig. S12). Many reports in the literature fail to disclose critical environmental details such as temperature or pH, which drastically limited our ability to make a cohesive body of data covering the environmental properties. The simulated body of data were unified as we homogenously ran DFT simulations and extracted identical parameters from the optimized results. Table S1 lists the parameters initially considered.

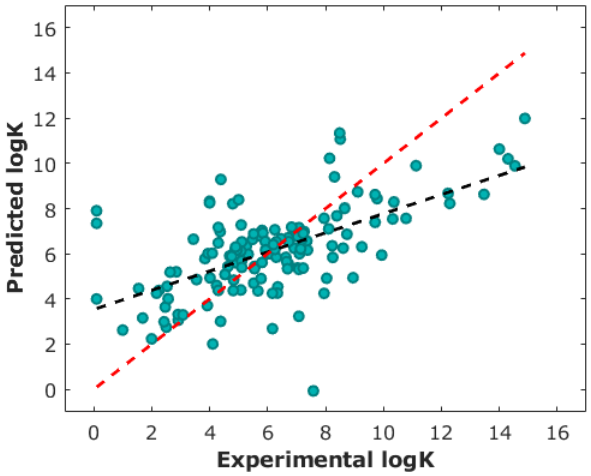


***Figure 1.*** Prediction performance for optimal leave-one-out experiments. The line of best fit based on predicted points is shown in black, and the line representing perfect prediction is shown in red (score = 1.50).

With 196 molecular samples consisting of 17 experimental and structural features, the constructed data set is small sample-wise with a relatively high-dimensional feature space. Without heavily sub-setting the feature space and losing potentially integral feature-interaction information, training a model to find a subspace parameterizing the underlying binding dynamics is difficult without strong inductive biases provided *a priori*. Without such biases, we instead looked to kernel methods, dual to more traditional subspace-learning methods, to provide a more sample-efficient learning paradigm that can still capture the dynamics of the feature space through the lens of properly-defined sample similarity. A mathematical background for kernel methods is provided in the supporting information.

Kernel featurization provides a non-linear representation of the samples within some inner-product space. Support Vector Machines (SVMs) are a family of models that can capitalize on this expressive kernel structure by representing examples as points in this space and determining an optimal but well-behaved mapping that best describes the differences between individual points. Although originally designed for classification tasks, SVMs have a natural extension to regression. Given the mathematical framework developed in the supplementary information, we explored the capacity of SVMs to predict the equilibrium binding constants of published data.[6] We performed a search over features to determine the best-performing subset of the feature space in coordination with grid search over hyperparameters within the model pipeline, namely $\gamma, \epsilon, C, |\theta|, \sigma$, and all permutations of addition or multiplication of each kernelized feature. The optimal hyperparameters (see supplementary information) were chosen based on 5-fold cross-validation.

The environmental data were largely incomplete due to the fact that many experiments in the literature do not report at least one and often several of the environmental parameters such as temperature or pH. For samples missing this information, we assumed temperatures of 298.15 K, and pH values of 7. We also set other values, such as salt concentration, to zero. These assumptions resulted in an environmental feature set that was sparse and largely uniform (see below). Viewing environmental factors as a single feature vector, we also explored how the addition of environmental information affected prediction performance.

A leave-one-out analysis was performed where the optimal model was trained on the entire training set less one sample. The log of the equilibrium constant, $\log K$, was then predicted by the model for the held-out sample. Mean absolute error was calculated across every combination of the 8 features listed in Table S1, and the subset with the lowest error was chosen to go forward (Fig. S13-16, S18, S19). Because the available environmental data lack diversity and are unnaturally uniform across samples, their usage as an additional feature often masked the underlying predictive capacity of the structural features. This process of feature reduction resulted in an optimal model consisting of 4 features derived from DFT calculations: optimized orientation, SCF density, electrostatic properties of each atom, and the overall electric field gradient (Fig. S18). These results are intuitive: both the size and electron distribution of small molecule organics are key in determining binding to cucurbit[7]uril.[6] Environmental parameters including salt concentration are known to affect the binding of some molecules.[6] However, the extent of changes is less than the error of our model, so environmental parameters were not considered going forward.

Optimized orientation was the largest driver of model accuracy in predicting $\log K$ (see supplementary information). In pursuit of better intuition regarding model performance, the equivalent SVM classifier was trained using the same process as above. The confusion matrix in Figure S19 is largely diagonal, with a bias towards over-predicting samples with a low value for $\log K$. Also of interest was the extent to which the preprocessing methods provided separation between samples. Figure S17 shows non-linear 2D projections of the combined kernels as well as the pre-kernelized and post-kernelized features for the optimized DFT orientation.[16] It is evident from these plots that the featurization process creates useful separation between high and low values of logK.

We next sought to challenge the model and identify its limits. Since the environmental data were disparate and incomplete, our final model did not use these data. We re-

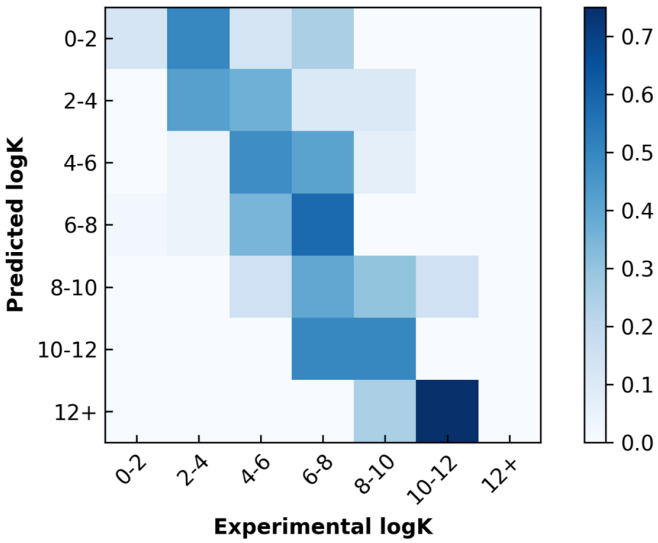


***Figure 2.*** Normalized confusion matrix for the optimal SVM classifier.

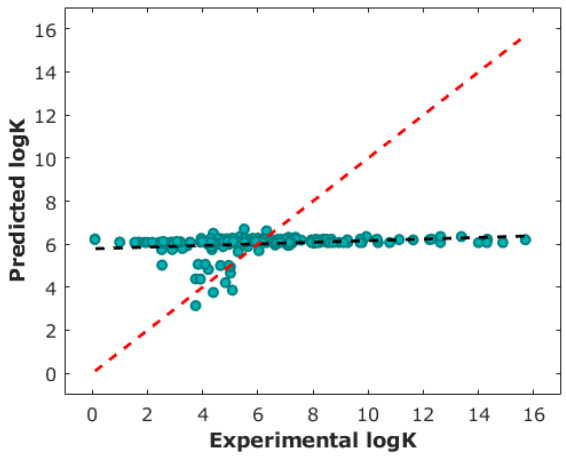


***Figure 3.*** Prediction performance for leave-one-out experiments for environmental parameters only. The line of best fit based on predicted points is shown in black, and the line representing perfect prediction is shown in red.

moved any duplicate molecules at different conditions and set the true logK as the average of all the reported values. For example, methyl viologen was reported 14 times at different parameters such as temperature or salt concentration, and so instead of having methyl viologen appear 14 times, it appeared once (ESI). Interestingly, and perhaps expectedly, the optimal model in this duplicate-free data set remained the same. The duplicate-free data set was chosen for subsequent analysis and the performance, confusion matrix, and corresponding ROC-AUC plot are reported (Fig. 1, 2, S17).

With a model in hand, we next sought to find its limitations. First, we removed classes of families and tested the model's ability to predict any one member of that family (Table 1, Table S2, Fig 4). For a family with $n$ members, a score was defined.

$$Score = \frac{1}{n}\sum_{i=1}^{n} |logK_{i,actual} - logK_{i,predicted}|$$

Given the limited size of the dataset, we expect this algorithm to be useful in identifying the binding of molecules which a supramolecular chemist might expect to bind to cucurbiturils *a priori*. For example, molecules with extended aromaticity are generally hypothesized to have some kind of activity with cucurbiturils.[15] This analysis shows that in order to capture binding of molecules such as imidazolium derivatives and adamantyl compounds, a data set containing these molecules is required. Imidazolium derivatives performed the best out of all the groups considered when their family was included in the training, and their error increases more than 5 times when left out, suggesting the algorithm is particularly sensitive to training on these types of molecules. Small arylamines and viologen derivatives performed better than the average data set regardless if the family was kept in or out, suggesting analysis of these kinds of molecules is robust and the physics of their binding is well-captured with the remaining data.

We next performed classical machine learning controls.[17] We first tested the performance of environmental parameters alone, which contain no chemical information about the guest. We found they had poor predictive capabilities (Fig. 3). We also tested whether we could predict the logK by counting the number of carbons in each molecule (Fig. S21). Similarly, we found poor predictive capabilities with this approach. As expected, both models performed worse than models which considered 3D structural data. One potential bias in the data that could be leading to the difference in the controls' performance is the slight negative relationship of molecular weights of guests (Fig. S22) to logK. Finally, we generated a random data set of identical dimension with the same logK outputs and found this had poor predictive capabilities (Fig. S23). We also randomly reassigned logK values to different input data and found this reshuffling had, as expected, poor predictive capabilities (Fig. S24).

Within the domain of utility, this model can provide an order-of-magnitude approximations of binding constants of molecules we might suspect *a priori* have binding to cucurbiturils. The mode can discriminate between molecules with no binding, moderate binding, and strong binding (Fig. 2). We next utilized it to predict whether binding can occur between cucurbit[7]uril and two small molecule organics recently identified as potentially promising drugs against pediatric low-grade gliomas: a type II RAF inhibitor TAK-580 (formerly MLN2480; referred to here as RAF), and a MEK inhibitor Selumetinib (also called AZD6244; referred to here as MEK).[18,19] Sun and colleagues recently reported RAF as a more promising therapy than type I RAF inhibitors due to its ability to bind to both fused and truncated v600.[18] Banerjee and colleagues also recently reported a promising phase I clinical trial of MEK in children with low-grade gliomas.[19] We performed DFT geometry optimizations on these two molecules and applied the SVM model. It was predicted that RAF would be a good guest to CB[7] with a logK of 4.61, while MEK would have very poor binding with a logK of 1.18 (Fig. 5C). Similar values were obtained if duplicate inputs were considered (Fig. S20). Synergistic drug cocktails have more potent responses than the sum of their individual components.[20] A key challenge in developing drug cocktails is in their delivery because drugs have different therapeutic windows requiring different release kinetics.[21,22] The ability to independently modulate release kinetics is an invaluable tool in the development of combination drugs. Different binding constants with macrocycles such as cucurbit[7]uril is one promising approach to independently modulate these kinetics. This prediction that two promising drugs (Fig. S25) against pediatric low grade gliomas is a potentially promising 'hit' in combination drug delivery.

We experimentally validated whether these predictions on the strong and poor CB[7] binding of RAF and MEK were accurate. Upon addition of CB[7] to an aqueous solution of

***Table 1.*** Summary of different subclasses of molecules identified in the data set that were used to challenge the model.

| *Family of molecules* | *Unique entries* |
|---|---|
| small arylamines | 4 |
| viologen derivatives | 6 |
| methylene blue derivatives | 9 |
| perfluorinated compounds | 13 |
| amino acids | 10 |
| imidazolium derivatives | 8 |
| adamantyl compounds | 12 |

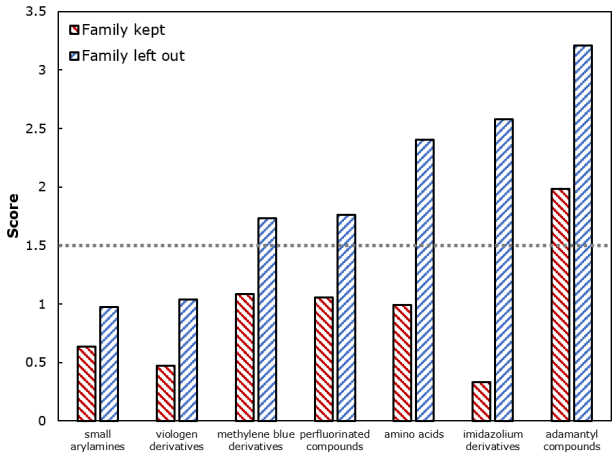


***Figure 4.*** The score describes the mean difference between predicted logK and actual logK when each class of families is kept or left out. Dashed grey line is the average score of the model utilizing all the data.

RAF (1:1 molar ratio), the drug's aromatic $^{1}$H NMR peaks remained sharp and well resolved. The proton signals of CB[7] split into two sets of equivalent peaks (Fig. 5A). These two observations strongly suggest that RAF and CB[7] bind favorably and statically.[15] We also sought to identify precisely where RAF was binding with CB[7]. No information on $^{1}$H or $^{13}$C NMR peak assignments could be found on RAF from the manufacturer or in the literature, and so further characterizations were carried out (Supplementary Section S.3: Binding Analyses *via* NMR). Our results show that CB[7] binds statically at the trifluoromethyl-substituted ring in a 1:1 fashion (Fig. S29). Surprised by this result, we sought to understand why CB[7] preferentially bound to the bulkier trifluoromethyl-substituted ring if there was an alternative pyrimidine with a positively charged amine.[6] Deuterated hydrochloric acid solution (0.1 M) was titrated into a solution of RAF alone (Fig. S30). The aromatic peak meta to the primary amine shifted after a reduction to pH $\leq$ 2. This suggests that the primary amine is, in fact, uncharged, which may be a reason why CB[7] does not bind at the pyrimidine ring. We then investigated whether RAF could bind to CB[8] (Fig. S31). The aromatic peaks of the drug do not remain well resolved as in the case with CB[7], but rather they broaden and disappear. This suggests that RAF does interact with CB[8] with low affinity and in a highly dynamic manner. Thus, CB[8] is not a good carrier for RAF, while CB[7] is an excellent one with $K_{CB[7]} = 3.5\times10^{6}$ M$^{-1}$ (Fig. S29).

We then validated whether the SVM prediction for MEK was correct. MEK was added to an aqueous solution of excess CB[7] to determine whether any interactions were occurring (Fig. S32). In depth analysis is described in the ESI. These data demonstrated that the drug does not bind to CB[7], confirming that the SVM predicted poor binding of MEK with CB[7]. We then screened its binding to CB[8] (Fig. S33). The shift and retention of sharp peaks in the $^{1}$H NMR spectra suggested that the MEK inhibitor binds more strongly and statically to CB[8] than RAF. The downfield shifts of protons *c*, *g*, and *h* suggested that the extended imidazole ring is located near but outside the CB[8] cavity. The upfield shift of protons *a* and *b* suggested that the ethylene glycol unit is inside the CB[8] cavity. The minimal changes in protons *d*, *e*, and *f* were consistent with the hypothesis that the bromo-substituted ring was not inside or near the CB[8] cavity. It is well known that CB[8] can thread poly(ethylene glycol) chains.[6] The thermodynamically favorable interactions between ethylene glycol repeat units and CB[8] may explain why CB[8] preferentially binds to the ethylene glycol unit of MEK. After addition of CB[8] in ratios greater than 1:1, little change occurs in the spectra, which suggested MEK and CB[8] bind in a 1:1 fashion. These data show that two different drugs with different therapeutic windows bind to different CB macrocycles. MEK shows no binding with CB[7], yet RAF and CB[7] bind strongly in a 1:1 fashion. Conversely, MEK binds to CB[8] more statically than RAF. Combining these two drugs into one therapy could give rise to a paradigm that provides a unique opportunity to selectively tune the release or residence time of one drug independently of the other by simply tuning the concentrations of CB[7] and CB[8] in the delivery system.

We next sought to provide a qualitative example of the potency of these drugs, and why modulating drugs to have different release kinetics is an important capability in the development of combination therapies. RAF/MEK combination therapies have been found to be efficacious against other malignancies including leukemia and colorectal cancers.[23] Recently BRAF and MEK dosages combined with PD-1 blockade was shown to be an effective immunotherapy approach against melanoma.[24] We hypothesized that such a combination may prove potent in a pediatric glioma model. We screened for combinations of RAF and MEK against a v600e mutant and identified a synergistic effect at $10^{2}$ nM concentration of both RAF and MEK together (Fig. S34). This result suggests that by co-delivering RAF and MEK, the concentration required of drug can be reduced at least 100 fold to achieve the same outcome. Further optimizations may yield further reductions in required concentrations.

Finally, we develop a model to showcase how with these binding affinities, CB[7] can be used to independently tune

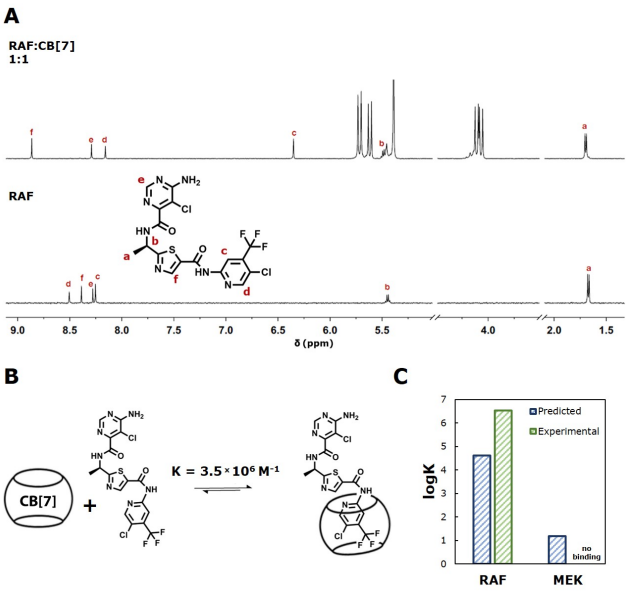


***Figure 5.*** (A) $^{1}$H NMR spectra of RAF alone (bottom) in DMSO-d6/$D_2O$ solution, and with CB[7] in a 1:1 molar ratio (top) in the same solution. (B) Illustration showing geometrically accurate binding of RAF with CB[7]. (C) Predicted and experimental logK of RAF and MEK.

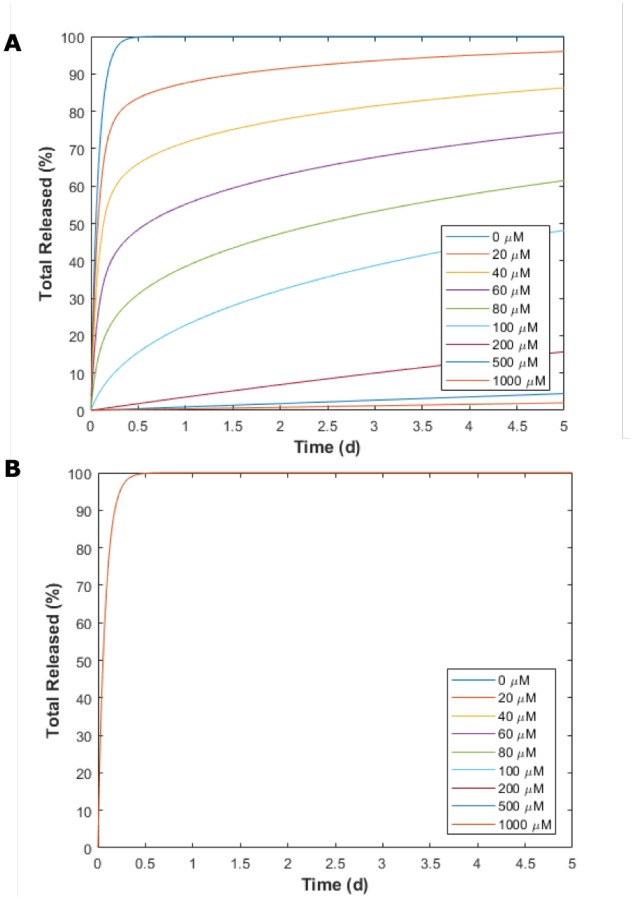


***Figure 6.*** Simulated time-resolved release kinetics of RAF and MEK with different concentrations of CB[7]. Figure shows five day result of (A) RAF and (B) MEK. MEK shows no change in concentrations as it does not bind to CB[7].

the release kinetics of one drug without changing the kinetics of the other (Fig. 6). We model a spherical, non-degradable hydrogel depot 0.375 mL in volume with bound CB[7] in the matrix and 100 $\mu$M loaded drug concentrations. The concentration of loaded CB[7] was varied. Across different concentrations of CB[7], the release kinetics of RAF changed several orders of magnitude in timescales (Fig. 6A, S35). By contrast, the changing concentration of CB[7] did not change the release kinetics of MEK (Fig. 6B). This result shows the utility of this macrocycle in tuning the kinetics: a high concentration can be loaded, and release can be prolonged over time scales of interest for local drug delivery.[21]

In this work, DFT calculations were used as training data to predict equilibrium binding constants of small molecule organics to CB[7] with machine learning. A library was developed and used to identify which parameters provide predictive capability. This algorithm was then used to predict the binding of two promising small molecule drugs in the clinic against pediatric low grade glioma. The algorithm predicted strong binding for the type II RAF inhibitor, and poor binding for the MEK inhibitor, which was experimentally validated. It was also discovered that CB[7] is partial to binding the RAF inhibitor, and CB[8] is partial to binding the MEK inhibitor, suggesting an opportunity for tunable release kinetics by introducing different concentrations of CB[7] or CB[8] into the system, perhaps in a hydrogel depot. Finally, we qualitatively demonstrated that these two drugs have different therapeutic windows and may have utility in concert against low grade gliomas. Machine learning may prove valuable in the development of drug delivery materials for combination therapies in the future, as well as non-biomedical applications that requires predicting the binding of small molecules to macrocycles. This work represents an original effort to bring machine learning to the field of supramolecular chemistry. As datasets continue to be generated and refined, the opportunities of data science in supramolecular chemistry will continue to grow.

**Acknowledgement** A.T. and M.P.S. thank The Winston Churchill Foundation of the United States. A.T. thanks the National Science Foundation graduate research fellowship, the MIT Chemical Engineering first year fellowship, and the Churchill College post-graduate grant program. G.W. thanks the Leverhulme Trust (project: 'Natural material innovation for sustainable living'). V.K.R. thanks the Swiss National Science Foundation (P2EZP2_168784). The authors thank Prof. Lucy Colwell (Cambridge) for fruitful discussions on machine learning controls, Prof. Charles Stiles (Harvard) for providing API stocks, Prof. Jeremy Baumberg (Cambridge) for support with DFT calculations, and Prof. Dane Wittrup (MIT) for support with kinetic modeling. The authors also thank Prof. Connor Coley (MIT), Dr. Magdalena Olesińska (Cambridge), Dr. Stefan Mommer (Cambridge), and Clement Hallou (Cambridge) for engaging and useful discussions.

## References

(1) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. *ACS Central Science* **2017**, *3*, 434–443.
(2) Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; Thrun, S. *Nature* **2017**, *542*, 115.
(3) Coley, C. W.; Green, W. H.; Jensen, K. F. *Accounts of Chemical Research* **2018**, *51, 5*.
(4) Hannun, A. Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G. H.; Bourn, C.; Turakhia, M. P.; Ng, A. Y. *Nature Medicine* **2019**, *25*, 65–69.
(5) Coley, C. W. et al. *Science* **2019**, *365*.
(6) Barrow, S. J.; Kasera, S.; Rowland, M. J.; Del Barrio, J.; Scherman, O. A. *Chem. Rev.* **2015**, *115*, 12320–12406.
(7) Palma, A.; Artelsmair, M.; Wu, G.; Lu, X.; Barrow, S. J.; Uddin, N.; Rosta, E.; Masson, E.; Scherman, O. A. *Angewandte Chemie - International Edition* **2017**, *56*, 15688–15692.
(8) Appel, E. A.; Rowland, M. J.; Loh, X. J.; Heywood, R. M.; Watts, C.; Scherman, O. A. *Chemical Communications* **2012**, *48*, 9843–9845.
(9) Kuok, K. I.; Li, S.; Wyman, I. W.; Wang, R. *Annals of the New York Academy of Sciences* **2017**, *1398*, 108–119.
(10) Werle, M.; Bernkop-Schnürch, A. *Amino Acids* **2006**, *30*, 351–367.
(11) Muddana, H. S.; Fenley, A. T.; Mobley, D. L.; Gilson, M. K. *Journal of Computer-Aided Molecular Design* **2014**,
(12) Ng, A. *deeplearning.ai* **2018**, *1*, 1–61.
(13) Frisch, M. J. et al. *Gaussian Inc.* **2009**, Wallingford CT.
(14) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *Journal of Physical Chemistry* **1994**, *98*, 11623–11627.
(15) Wu, G.; Olesińska, M.; Wu, Y.; Matak-Vinkovic, D.; Scherman, O. A. *J. Am. Chem. Soc.* **2017**, *139*, 3202–3208.
(16) Maaten, L. v. d.; Hinton, G. *Journal of machine learning research* **2008**, *9*, 2579–2605.
(17) Chuang, K. V.; Keiser, M. J. *Science* **2018**, *362*.
(18) Sun, Y. et al. *Neuro-Oncology* **2017**, *19*, 774–785.
(19) Banerjee, A. et al. *Neuro-Oncology* **2017**,
(20) Sugahara, K. N. *Science* **2011**, *328*, 1031–7.
(21) Tabet, A.; Jensen, M. P.; Parkins, C. C.; Patil, P. G.; Watts, C.; Scherman, O. A. *Advanced Healthcare Materials 8*, 1801391.
(22) Tabet, A.; Wang, C. *Advanced Healthcare Materials 8*, 1800908.
(23) Gibney, G.; Messina, J.; Fedorenko, I.; Sondak, V.; Smalley, K. *Nat Rev Clin Oncol* **2013**, *10*, 390–399.
(24) Ribas, A. et al. *Nature Medicine* **2019**, *25*, 936–940.