Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d6sc00479b

All publication charges for this article have been paid for by the Royal Society of Chemistry

Synthesis and machine learning techniques to enable data-driven investigation of supramolecular host–guest interactions

Alok Shaurya,†[a] Amir Hassan Bagherzadeh Mostaghimi,†[b] David R. Turnbull,[b] Fraser Hof *[ac] and Jeffrey F. Van Humbeck *[b]

Received 16th January 2026
Accepted 4th May 2026

DOI: 10.1039/d6sc00479b

rsc.li/chemical-science

The availability of large datasets such as the Protein DataBank and ChEMBL have allowed for rapid progress in developing machine learning tools for predicting the biological activity of organic small molecules. The binding between supramolecular hosts and their desired guests is governed by the same forces that drive protein-small molecule interactions, and yet this field has seen dramatically less application of machine learning. In this contribution, we demonstrate that the production of easily diversified building blocks can allow a single laboratory to generate a dataset that is sufficient to engage with modern machine learning approaches. A range of methods were evaluated against our single-laboratory dataset, with a graph neural network featuring an attention mechanism providing meaningful performance in this data-sparse arena.

Introduction

Organic chemistry has seen an accelerating adoption of machine learning (ML) and other data science tools.[1–6] Two complementary sources of data have played significant roles in this evolution. First, there are the large, well-curated, open-source repositories such as the Protein DataBank[7] or European Molecular Biology Laboratory-European Bioinformatics Institute's (EMBL-EBI) ChEMBL small-molecule database.[8] Second, well-resourced industrial and academic research groups have made great strides in developing high-throughput techniques in synthetic organic chemistry, allowing for hundreds or thousands of reactions to be carried out.[9–15] These latter investigations have often provided a double benefit, yielding both highly optimized systems for a specific task and mechanistic understanding that illuminates synthetic organic chemistry more broadly.

Supramolecular chemistry faces significant challenges when attempting to apply the same approach. There are no databases of comparable size to those that are widely used for biological or small-molecule chemistry.[16] Research groups with ML expertise to inform the community, therefore, are unable to demonstrate which specific approaches might be useful. In the area of host–guest chemistry, there is a significant problem of publication bias in the literature. Most publications focus on the extraordinary: unique systems that show dramatic binding affinity or selectivity between competitive substrates. The synthesis of molecular hosts that display these properties has been similarly goal-oriented: whether enough of the singular target host is made is often the only concern. Parallel synthesis of macrocyclic hosts has not been a traditional pursuit of the discipline and has involved relatively low-throughput efforts.[17] Herein, we demonstrate that synthetic innovation can allow for a single experimental research group to generate sufficient molecular libraries for data-driven investigations. Realizing that experimental research groups might possess varying levels of comfort with machine learning techniques, we analyzed our single-lab dataset with five different approaches. These approaches were chosen to span from simple, widely applied tools that can be implemented with limited experience, to more powerful models. We aimed to predict binding affinity (*i.e.* regression), to complement previous work on classification,[18] and host generation,[19] where there have been successful reports. A recent preprint combines ML with GFN2-xTB calculations to estimate binding free energy of small anion guests.[20]

Results and discussion

Host and guest target selection

Protein methylation is a common regulatory mechanism in higher organisms. On certain amino acid side chains, the number of methyl groups (Fig. 1) is a functionally unique post-translational modification.[21–25] As the distinct roles of different lysine and arginine methylation marks have become better

[a]*Department of Chemistry, University of Victoria, 3800 Finnerty Rolad, Victoria, BC V8P 5C2, Canada. E-mail: fhof@uvic.ca*

[b]*Department of Chemistry, University of Calgary, 2500 University Drive, Calgary, AB T2N 1N4, Canada. E-mail: jeffrey.vanhumbec1@ucalgary.ca*

[c]*Centre for Advanced Materials and Related Technology (CAMTEC), University of Victoria, 3800 Finnerty Road, Victoria, BC V8P 5C2, Canada*

† These authors contributed equally.

understood, the need for new binding agents that specifically target different marks has arisen. In response to the identification of these targets, strong recognition of trimethyllysine by macrocyclic hosts has been routinely achieved, including by sulfonated calix[4]arenes,[17,26–30] dithiamacrocycles,[31,32] cavitands,[33,34] and more.[35,36] The search for agents that selectively bind lower methylation states of lysine or specific methylation states of arginine represents a major challenge in biosupramolecular chemistry. Waters' group has been successful with a variety of dynamic combinatorial chemistry approaches. They were able to identify the first synthetic host that binds Kme2 selectively over Kme3,[37] and later a host that selected Rme2a over Rme2s.[38] An alternate strategy has been used when designing neutral hosts. Such hosts, like cucurbiturils[39] and phosphonated receptors,[40,41] substitute electrostatic interaction with ion–dipole interactions. In both examples, multiple dipoles, P=O or C=O, are pointed inward towards the cavity and surround the incoming guest.

Our own work towards selective binding of methylated amino acids has focused on calixarene frameworks. Arylating the upper rim extends the hydrophobic surface and gives hosts that bind Kme3 with low micromolar affinities.[27,29] A lower-rim substitution that confers selectivity for dimethyllysine was discovered accidently.[42] In addition to altering the binding surface, substitution on calix[4]arenes also alters their conformational preferences in unpredictable ways which have direct (and also hard to predict) influence on the binding properties.[28] Given the ability to generate a diverse library of these scaffolds (*vide infra*), we selected this host–guest pair as a reasonable target to test a data-driven approach around.

We aim for this work to complement the 'SAMPL' host–guest challenge, which provides an experimental binding dataset to benchmark computational approaches.[43] In the most recent competition (SAMPL-9), the magnitude of this challenge remains clear. For the most similar SAMPL host to our own (*i.e.* a pillararene derivative), root-mean squared errors ranging from 2.04–3.75 kcal mol^{-1} were reported in ranked submissions. The guests in this challenge are small (*e.g.* cyclohexylamine, paraquat) and feature much less conformational flexibility that the peptide guests investigated here. At the outset, we did not consider fully *in silico* approaches such as those reported in SAMPL as appropriate for our application.

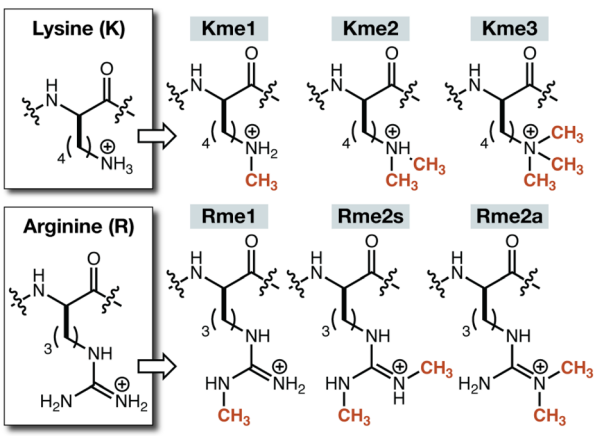


Fig. 1 Post-translational methylation of amino acid residues of relevance to this project.

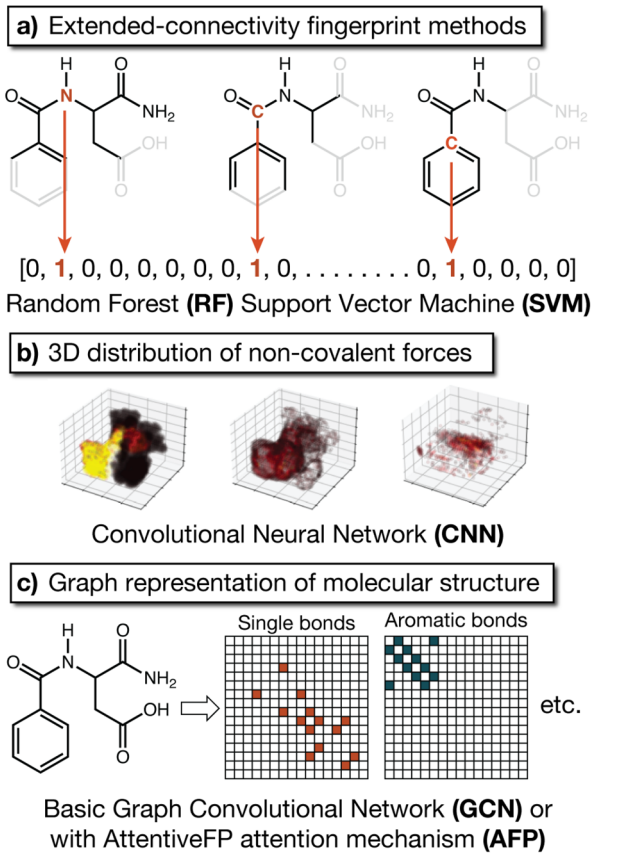


Fig. 2 An overview of the five machine learning approaches evaluated against our host–guest binding dataset. (a) ECFP methods leverage traditional machine learning tools and operate on a binary barcode of uniform length. (b) 3D modelling of non-covalent forces are voxelized to create a standard input for convolutional neural network analysis. (c) The adjacency matrix of a molecule is one representation of a graph approach, which can be analyzed with or without an 'attention mechanism'.

ML algorithm selection

Dozens of different ML approaches that try to predict the binding affinity of a small molecule to its protein target have been developed,[44–46] based in no small part on the data availability provided by the ChEMBL database. We chose to test 5 specific models that show 3 contrasting approaches to the problem, and demand different levels of skill in the art (Fig. 2). One of the simplest methods involves the use of Extended Connectivity Fingerprints (**ECFPs**; Fig. 2a).[47] In this approach, a molecular structure is converted into a binary bar-code of standard length. The local environment around each atom—out to a pre-determined radius—determines which positions in the bar code are assigned a value of one. While simple, this encoding provided competitive performance in an evaluation of over 5000 specific datasets extracted from ChEMBL.[48] We chose

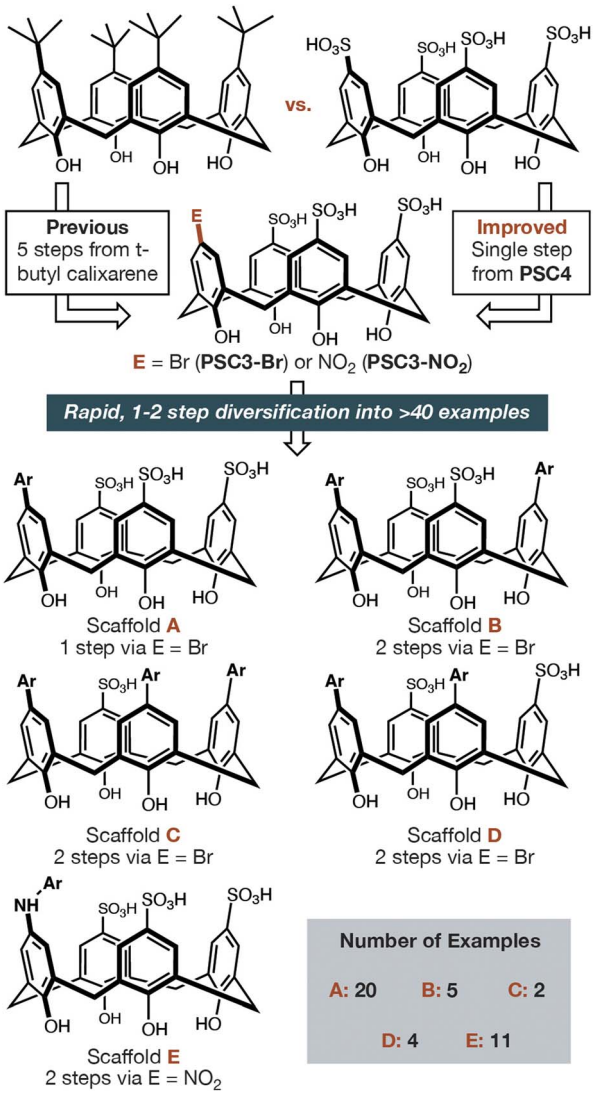


Fig. 3 Synthetic approach to a focused set of upper rim-substituted calixarene hosts, which yields 42 targets.

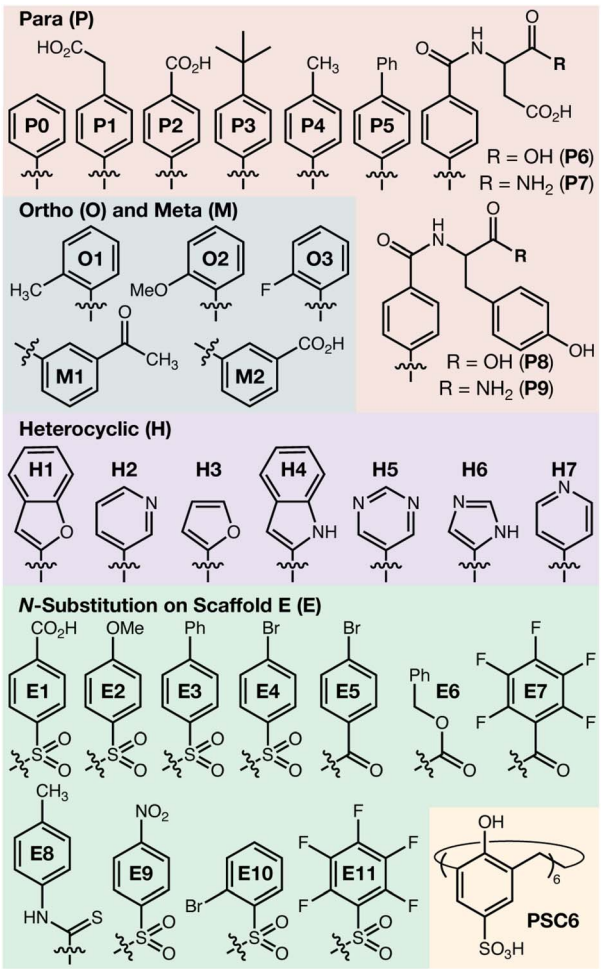


Fig. 4 Specific aryl substituents investigated in this work along with their corresponding library labels.

two specific algorithms—random forest (**RF**) and support vector machine (**SVM**)—as user friendly options that work with **ECFP**s. A lab with absolutely no coding experience could likely begin using such an approach within several weeks. The creation of **ECFP**s relies on the *rdkit* python package, and the implementation of **RF** and **SVM** use *scikit-learn*. These packages have extensive support and are easily accessible.[49,50]

To contrast **ECFP**s, we also investigated two approaches that represent molecules with more physics-informed data structures and have been analyzed extensively using ChEMBL. In one approach, the supramolecular host was modelled in 3D-space using standard tools (see SI). The distribution of steric occupancy, partial charge, and polarizable surfaces were recorded on a spatial grid which became the input for a convolutional neural network (**CNN**), in line with several existing approaches.[51–55] Next, we also considered the graph representation of our hosts. Such a framework represents molecules as a combination of nodes (*i.e.* atoms) and edges (*i.e.* bonds) through which information is shared. Several different ways exist to draw information out of the individual atoms to represent a given molecule as a whole. A basic graph convolutional network (**GCN**) was evaluated here.[56] As compared to the relatively easy **RF** and **SVM**, we would rate these next two approaches as being of medium difficulty to implement. Some basic coding proficiency is needed, though the widespread use of **CNN** and **GCN** in several different scientific areas means there is a great deal of available support.

A suite of more advanced techniques exist that aim to increase the performance of ML models in this domain. Finally, we selected a specific academic model—the 'AttentiveFP' approach of Jiang and Zheng—for inclusion.[57] AttentiveFP also uses a graph representation of molecules but adds additional facets to the approach to learn which sub-structures of a given molecule the algorithm should pay attention to, hence the term 'attention mechanism'. The original work has been cited several hundred times and affords a balance between being well-established while also showing what is possible at the leading edge of academic work. We were able to successfully use the resources provided by those groups to evaluate our own dataset

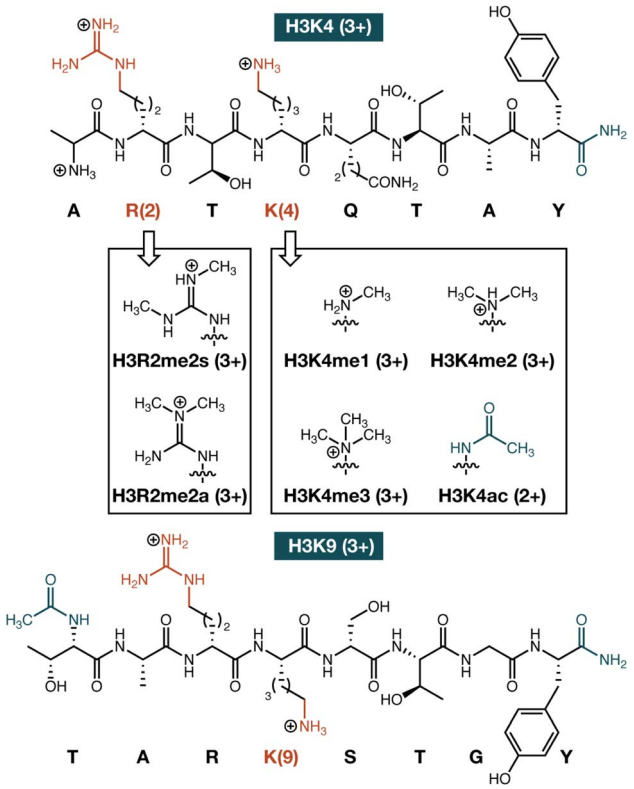


Fig. 5 Structures of all peptide guests evaluated.

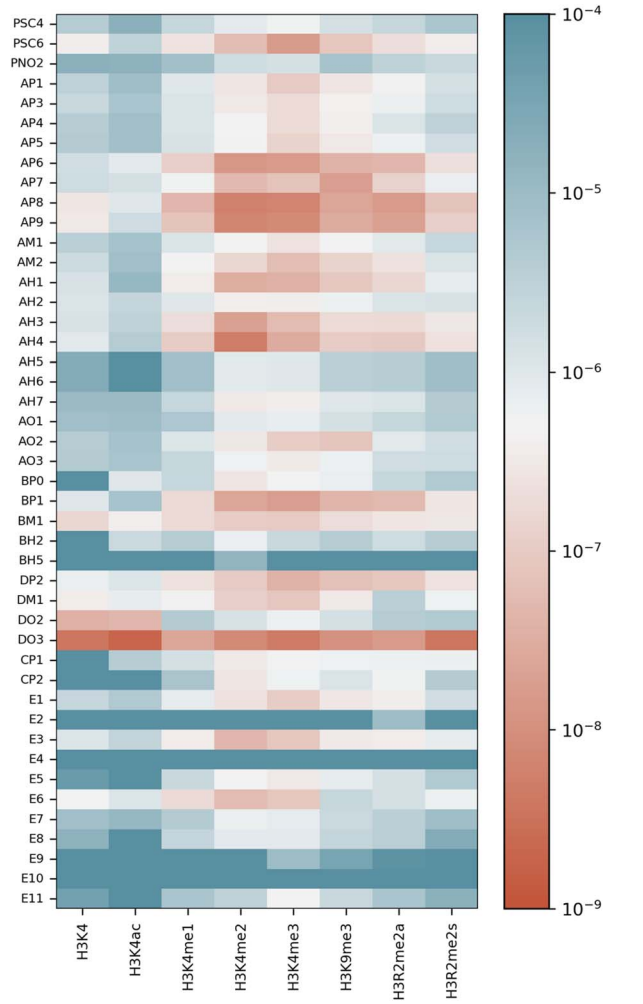


Fig. 6 Binding affinities between calixarene hosts and peptide guests as evaluated by fluorescence displacement assay. Color labels are of K_d as indicated in the legend.

(*i.e. via* their GitHub page), but recognize that this task might be difficult for a new entrant to the area.

Synthetic library generation

Our successful plan for rapid generation of a library of calix[4]arenes was executed as shown in Fig. 3. While literature based methods for regioselective synthesis of upper-rim substituted structures typically involves 5 (or more) synthetic steps to introduce a diversifiable functional group handle in the position of interest, we thought that direct desulfonylative substitution should be possible.[58,59] Beginning with commercially available **PSC4**, addition of sub-stoichiometric quantities of NBS delivered 35% yield of the mono-brominated derivative **PSC3-Br**, along with 50% recovered starting material. In all attempts to react **PSC4** with other sources of bromine (*i.e.* Br_2, or mixtures of bromide salts and oxidants) we saw no successful production of the desired **PSC3-Br**. Similarly, there existed precedent in the literature for the direct installation of an aromatic nitro group, at the expense of a sulfonic acid.[60] Based on a literature report, we tested combinations of protic acids and sodium nitrite and found dilute HCl an acceptable choice. A similar result to bromination was achieved, where 35% yield of **PSC3-NO$_2$** could be obtained along with 40% recovered starting material.

Initial efforts towards the mono-bromination reaction had shown that multiply-brominated byproducts could be produced. Some of these had molecular masses and isotope patterns consistent with multiple brominations. While we were never able to discover reaction conditions that would cleanly convert starting **PSC4** into di- or tribrominated scaffolds, we found that exposing purified **PSC3-Br** to our standard conditions generated a mixture of both possible disubstituted isomers as well as trisubstituted material. All of these isomers could be separated by HPLC in 44% combined yield with 40% unreacted **PSC3-Br** being returned.

With rapid routes to five different calixarene scaffolds established, we completed the synthesis of more than 40 derivatives with minimal additional synthetic steps (Fig. 3: scaffolds **A–E**). Any of the mono-, di-, or tribromo scaffolds could be elaborated *via* Suzuki coupling. Our optimized conditions are compatible with a range of simple arylboronic acids as well as more challenging heteroaryl and *ortho*-substituted coupling partners. 31 scaffolds were generated in this way, with mono-substitution favored (20/31) mostly for the high aqueous solubility of host molecules that retain three sulfonic acid groups. From the nitro-substituted scaffold **PSC3-NO$_2$**, reduction to the air-sensitive aniline followed by reaction with various sulfonyl or (thio)carbonyl electrophiles delivered

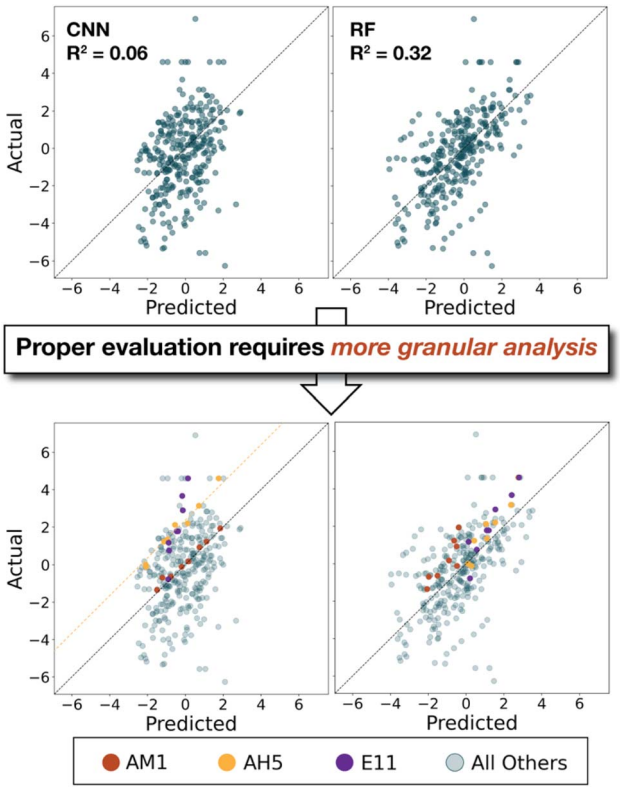
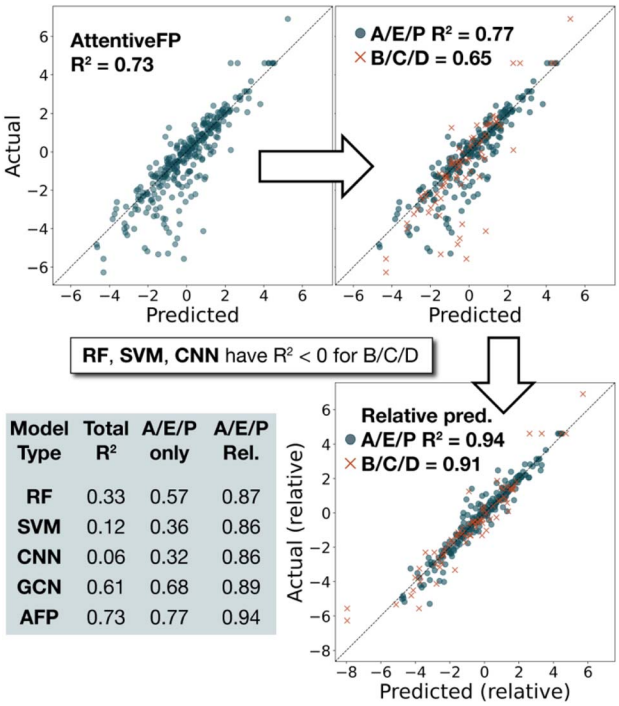


Fig. 7 Predicted *vs.* actual plots for **CNN** and **RF** models. Scales are measurement of ln(K_d). Evaluation of predictions for single calixarene hosts do not necessarily match the overall trend observed.



Model Type	Total R^2	A/E/P only	A/E/P Rel.
RF	0.33	0.57	0.87
SVM	0.12	0.36	0.86
CNN	0.06	0.32	0.86
GCN	0.61	0.68	0.89
AFP	0.73	0.77	0.94

Fig. 8 Different views of the dataset that became instructive in this work. Consideration of mono- or unsubstituted calixarenes *vs.* multi-substituted showed clear differences. Considering the predition of absolute affinity between one host and one peptide guest, *versus* predicting the relative affinity of two peptide guests (*i.e.* ln[K_{d1}/K_{d2}]) for the same host also showed significant differences between models (*vide infra*).

an additional 11 hosts. The specific substituents added are enumerated in Fig. 4. For the discussion below, the combination of a scaffold letter is followed by the substituent label to describe any given calixarene. The contents of the library are as follows: Scaffold **A** was elaborated with 8 *para*-substituted aromatics (**AP1**, and **3–9**), 2 *meta*-substituted (**AM1**, **2**), 3 *ortho*-substituted (**AO1-3**), and 7 heterocycles (**AH1-7**). Scaffold **B** (**BP0**, **BP1**, **BM1**, **BH2**, **BH5**), **C** (**CP1**, **CP2**), and **D** (**DP2**, **DM1**, **DO2**, **DO3**) were constructed with a small sub-set of those shown. All 11 'E'-labelled structures were connected through the aniline nitrogen on scaffold **E** to generate **E1–11**. Finally, the larger calixarene homologue **PSC6** was also procured for binding studies. Including the starting material **PSC4** and intermediate **PSC3-NO$_2$** provided 45 hosts for binding studies.

Host–guest association

To determine the diversity of this library's binding capabilities, we chose a set of post-translationally modified guests in which a relatively small change in modification is embedded within a larger (common) peptide structure (Fig. 5). All guests used in this study are peptides that are eight amino acids long, with the base sequence derived from the N-terminal tail of histone 3 (H3). The eighth position is tyrosine in all cases in order to aid UV detection during HPLC purification of the peptides. Each peptide has one specific post-translational modification, the identity and position of which is mentioned in the name.

Model Type	Training Style	Training Data	A/E/P R^2	B/C/D R^2	A/E/P Median	A/E/P Rel. R^2
RF	Abs	**All**	**0.57**	**-0.12**	**0.49**	**0.87**
		A/E/P	0.68	n/a	0.52	0.88
	Rel	All	0.44	-0.20	-0.07	0.89
		A/E/P	0.70	n/a	0.55	0.91
CNN	Abs	**All**	**0.32**	**-0.43**	**-0.28**	**0.86**
		A/E/P	0.22	n/a	-0.13	0.82
	Rel	All	0.46	-0.32	-0.09	0.88
		A/E/P	0.44	n/a	0.07	0.91
AFP	Abs	**All**	**0.77**	**0.65**	**0.82**	**0.94**
		A/E/P	0.83	n/a	0.77	0.93
	Rel	All	0.37	0.00	0.16	0.91
		A/E/P	0.37	n/a	-0.06	0.92

Fig. 9 Analysis of different training styles, and different dataset composition for three representative model types.

H3K4me2 for example is an 8-mer peptide made up of the first seven amino acids from the H3 tail (H3(1–7)). It has a tyrosine (Y) at its last position and lysine 4 is *N*,*N*-dimethylated (hence K4me2). All peptides had their N-terminus free (as in the native protein) and C-terminal amides. The exception is H3K9me3 peptide whose base sequence is H3(6–12) and has its N-terminus acetylated.

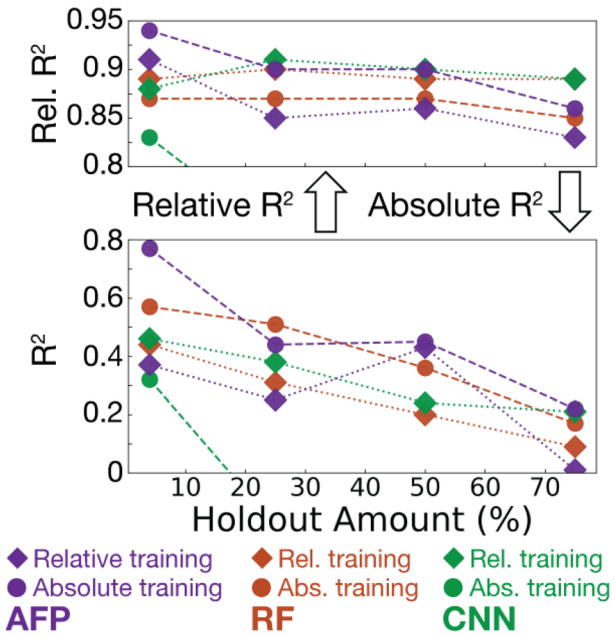


Fig. 10 Impact of training set size on model performance.

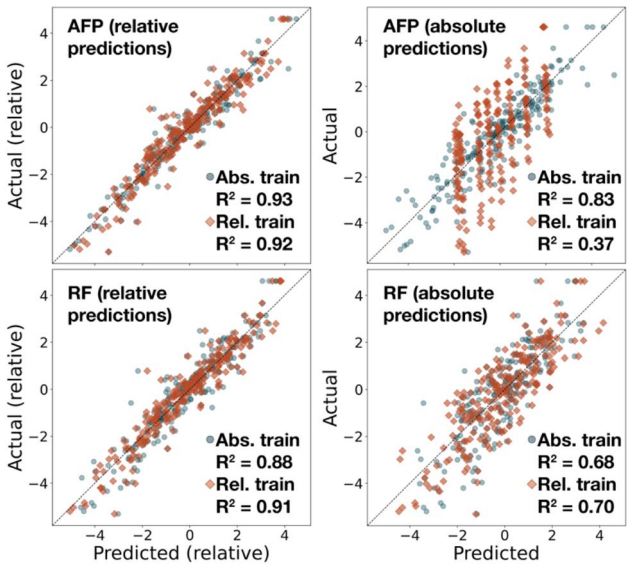


Fig. 11 Collapse of the AFP model to predicting roughly the mean value for each peptide host is observed when training on relative (K_{d1}/K_{d2}) affinity values. Such collapse is not observed for the RF model, for example.

Indicator displacement assays (IDAs) provided rapid access to a complete understanding of each host's binding selectivity among a panel of post-translationally modified peptides.[29] All calix[4]arenes in the library bind lucigenin reversibly[61] and quench its fluorescence, meaning that a single indicator can be used for all combinations of host and guest. Direct titrations of host into dye provided K_{ind}, and competitive titrations of each peptide into host–dye complex provided the guest-binding constant K_d. A 96-well plate was laid out in order to accommodate duplicate 5- or 6-point titrations for every host–guest combination (18 total titrations per plate), and control wells (see SI). All binding data are summarized in a heat plot in Fig. 6.

Given the subtle differences in guest structure, a major question is whether this approach would create and identify hosts with significant differences in binding affinities and selectivities. The IDA data shows binding behavior that allows us to identify some qualitatively interesting trends that support the need for further analysis. Scaffold **A** is a consistent performer. Every scaffold **A** member binds H3K4me3 with $K_d < 1$ μM. This is stronger than many proteins that bind Kme3 peptides *in vivo*.[62–65] The selectivity for H3K4me3 over H3K4 is also very uniform among scaffold **A** members with most compounds being 20–30 fold selective for H3K4me3. **AP1** has the highest selectivity (>100 fold). Scaffold **E** hosts produced more diverse behaviors than scaffold **A**. In general, all sulfonamide containing hosts lost the affinity to N-methylated peptides with only **E1** and **E11** retaining meaningful affinity. Hosts **E4**/**E5** and **E11**/**E7** are two matched pairs in our library where the same aryl group is attached *via* sulfonamide (**E4** and **E11**) and amide (**E5** and **E7**) linkage. In both cases, the sulfonamide containing member appears to be a broadly worse host, to the point where **E4** doesn't bind anything at all. We attribute this result to the higher flexibility of sulfonamide linkages, which allow the aryl group to fold in toward the cavity of calix[4]arene. The behaviors of **E6** and **E8**, which have a carbamate and a thiourea linkage, respectively, also corroborates this notion. The more flexible benzyl carbamate group of **E6** can bend inwards and completely disrupts binding, while the tolyl thiourea group of **E8** has geometric preferences similar to an amide bond, and thus retains binding. Evaluation of 9 compounds generated from scaffolds **B** and **D** revealed the differences between regioisomeric hosts. The most intriguing member from this subset is **DP0**, which could not be analyzed by our standard IDA as it completely lost the ability to bind lucigenin (a first in our experience) while its isomer, **BP0**, binds lucigenin as expected. The monosubstituted analog **AP0** (not included in this study) also binds lucigenin and has been reported elsewhere to bind methylated peptides.[29] This result becomes even more outstanding when we note that **DO2** and **DO3** show the expected binding to lucigenin. These compounds differ from **DP0** only in terms of a single *ortho* substitution at each appended aryl ring. In a complete reversal, **BM1** and **DM1** have nearly identical binding profiles to each other. While it is important not to draw broad conclusions from this limited set of data, the chaotic behavior of this subset of molecules reinforces our understanding that the substitutions affect the binding behavior of calix[4]arene in seemingly unpredictable ways that make rational design difficult.

Affinity prediction by ML

Of the host structures measured in Fig. 6, a handful were excluded from further analysis. Those that could not be accurately evaluated due to extremely weak binding (*i.e.* **BH5**, **E2**, **4**, **9** and **10**) were excluded. Given its structural dissimilarity to all

other frameworks, **PSC6** was also not included our ML analysis. Last, our procedure for generating an ensemble of conformers used the ANI-2x potential,[66,67] which is not compatible with bromine. To be consistent across all models, **E5** was therefore also excluded. With 38 remaining hosts—each measured with 8 peptide guests—our dataset contained 304 binding constants. After routine hyperparameter optimization (see SI), we conducted an initial top-level analysis of each network type using leave-one-out cross-validation (LOO-CV). To mimic how such systems would function when evaluating a new calixarene host, the 'one' in leave-one-out refers to the calixarene host itself: all 8 individual data points involving that host are held out together. Using simple R^2 as an initial metric, we saw a range of performance, with some networks having essentially zero predictive performance (**SVM**: 0.12, **CNN**: 0.06), graph-based models showing decent to good results (**GCN**: 0.61, **AFP**: 0.73), and the final approach at an intermediate value (**RF**: 0.33).

Deeper analysis, however, showed that these overall metrics might not be providing the most accurate analysis in all cases. As shown in Fig. 7, when considering all predicted data points pooled together, it is true that the **CNN** model is essentially a vertical stripe (*i.e.* zero predictive power), while the slightly better performing **RF** is beginning to bend onto the ideal actual *vs.* predicted line on the diagonal. However, a different picture can emerge when individual hosts are considered separately. For the **CNN**—while the overall model appears to have no predictive power—**AM1** is predicted almost perfectly—better than by the **RF**. Considering **AH5**, it is true the **CNN** has significant systematic error. However, if one is most concerned as we are with binding selectivity, rather than absolute affinity, **CNN** again predicts the key behaviors of **AH5** almost perfectly, as these points lie on a diagonal line parallel to the ideal. Neither **RF** nor **CNN** predict the behavior of **E11** with enough accuracy to be useful and the difference between 'bad' and 'truly awful', respectively, has an impact on the overall R^2 to the detriment of the **CNN**.

Therefore, we evaluated these models across several metrics, with a workflow of our thinking beginning with Fig. 8. When inspecting the results for each host individually, it became immediately obvious that all model types performed better on mono- or unsubstituted structures (*i.e.* scaffolds **A** and **E**, plus **PSC4** and **PSC3-NO$_2$**). In fact, only **GCN** and **AFP** had any predictive power for the 10 calixarenes selected from scaffolds **B**, **C**, and **D**. **GCN** and **AFP** typically show the worst outliers coming from these scaffolds as well, though the overall R^2 is improved to a smaller extent by considering the two groups separately. Last, we wanted to account for the observation discussed in Fig. 7, where accurate trends in binding across peptide hosts was observed for some calixarenes, even if the absolute magnitude of the predictions was off. By considering relative binding (*i.e.* $\ln[K_{d1}/K_{d2}]$), any trend predicted accurately would now fall on the ideal diagonal of an actual (relative) *vs.* predicted (relative) plot. **AFP** is still superior by any considered metric, and this final transformation to relative affinity must be pursued with some caution, for reasons described below.

Model evaluations

Realizing that the complexity of **AFP** might be beyond the capacity of new entrants to the field, we have tried to further specify the trade-offs that can be made between this high-powered model, and other simpler choices. For the discussion below, **RF** and **CNN** are detailed due to their contrasting behavior (*vide infra*), with corresponds analysis for **SVM** and **GCN** in the SI. We could think of two ways in which the performance of simpler models could be improved. First, given the improvement seen in **RF** and **CNN** when only the **A**/**E**/**P** scaffolds were considered, we took this idea one step further and trained/evaluated all models on only this focused dataset (again by LOO-CV). Second, we also considered that many applications for supramolecular systems are most concerned with selectivity between substrates. That is, the relative affinity of two different guests for a single host K_{d1}/K_{d2}. In our own work with these methylated peptides, one of our goals has been to develop chromatographic adsorbents that can improve the selectivity between post-translationally modified proteins.[42] So, we also trained models of all types to make these 'relative' predictions directly. With some straightforward math, networks trained to make such relative predictions can still evaluate absolute binding affinity. We wondered whether the increase in the size of the training dataset could improve the accuracy of simple models like **RF**.

A summary of the results are shown in Fig. 9. When considering these two approaches—focusing the training data and/or explicitly targeting relative affinity—the three network types each show distinct behavior. **RF** is most sensitive to the homogeneity of the training data (Fig. 9. Green *vs.* white rows in **RF** category). When scaffolds **B**/**C**/**D** are excluded, every relevant performance metric is improved with little sensitivity to whether absolute or relative affinity values were used for training. **CNN** displays the opposite behavior, where the composition of the dataset has little impact on performance, but training on relative affinity values improves both predictions of relative and absolute affinity (Fig. 9. White *vs.* grey rows in **CNN**). **AFP** displays a third, orthogonal behavior. Neither of these modifications improves performance in any meaningful way, with training on relative affinity values proving clearly destructive (Fig. 9. Orange *vs.* white rows in **AFP**).

A final evaluation of the robustness of these models was performed by holding out increasing amounts of data from the training set. When training on the full set that includes **B**/**C**/**D**, LOO-CV is equivalent to ~3% holdout. We evaluated the impact on performance (both absolute and relative R^2) from holding out 25%, 50%, and 75% of calixarenes from the training set. The results are shown in Fig. 10, and broadly reinforce the conclusions of Fig. 9. **CNN** performance was not only significantly improved by training the network on relative performance, such training also allowed for performance to decline less dramatically as more data was held out from training. This analysis also appeared to delimit the amount of data that is necessary for **AFP** to maintain its significant advantage. In terms of both absolute and relative predictions, simpler models become competitive when less training data is available. Given the very small size of

our starting dataset, it is not at all surprising that performance would decline to near-zero when most of the hosts have been removed: the final 75% split would leave only a handful of hosts (9) for training. What is perhaps more surprising is the degree to which predictions of relative binding affinity persist across the different splits. This can most likely be attributed to one particular feature of the affinity dataset (*vide infra*).

Conclusions

For researchers in supramolecular chemistry who would like to begin applying ML techniques to their own datasets, we believe that our results suggest there are two paths forward. For those groups with limited coding ability or limited access to GPU resources, ECFP fingerprint models can demonstrate some meaningful performance. In our specific case a random forest model proved to be more accurate, though we would encourage others to investigate multiple models as the results on our dataset were the opposite of what had been previously observe in a large survey of ChEMBL.[48] While this approach is within reach even for groups that are completely new to the area, it did prove in our hands to be very sensitive to the construction of the dataset. While we were able to see by close inspection that A/E/P-type calixarenes worked together best as a training set, this was not necessarily obvious *a priori*. With limited resources, a group setting off to generate their own library of novel synthetic hosts might need every example to be an appropriate target.

AttentiveFP demonstrated impressive predictive performance across all calixarene scaffolds, especially when considering the small size of the total dataset and the diversity of hosts. Beyond the necessity for more advanced skills, there was a second observed downside to this model. The predictive performance fell off more steeply as larger amounts of data were withheld from training, suggesting that our full dataset existed near the edge of applicability for this approach. Close inspection of Fig. 9 also has a cautionary result for AFP: whereas CNN and RF showed better performance under some conditions when relative affinity was used as the training target, AFP showed strange behavior. While the prediction of relative affinity was largely unchanged, the prediction of absolute affinity diminished dramatically. Looking at the predicted *vs.* actual plots (Fig. 11) immediately explains this behavior, as well as the observation from Fig. 10 that even very small amounts of training data allow for acceptable prediction of relative affinity.

This can be best explained by considering the vertical stripes of orange diamonds in the top right plot of Fig. 11. In effect, when trained on relative binding constants, the **AFP** model predicted nearly the mean K_d for each of the 8 peptides, learning only small adjustments to these mean values for each host structure. Given that even less effective models could predict relative affinity with R^2 – 0.85, we reason that this is about the success one can achieve by quickly learning the mean binding constant for each peptide while mostly ignoring host structure. When directly training a ML tool to predict relative affinity, we would strongly encourage others to calculate performance based against a 'null' model, such as the mean absolute affinity value for each different analyte. For the system studied here, using the per-peptide mean delivers null models with $R^2 = 0.29$ (all training data; **A**/**E**/**P** prediction only) and 0.42 (**A**/**E**/**P** only training) for predictions of absolute affinity, with structure essentially identical to that shown in the top-right plot of Fig. 11 (see SI: Fig. SI-69 and 70.)

In our eyes, many molecular host constructs should be as amenable to diversity-oriented synthetic innovation as the calixarenes were in our case. Parallel investigation of several guests geometrically increases the number of data points for machine learning, and also provides the opportunity to consider (or, explicitly train for) selectivity between guests. While there was not consistent behavior observed across ML models, the fact that a very simple model could deliver useful predictions in certain cases and an advanced model could deliver good results across the entire dataset is encouraging. The results here do speak to caution being necessary in dataset design. The highest performing **AFP** model was relatively robust to including unusual calixarene types during training, but performance degraded quickly as the amount of training data decreased. Simpler models likely have some utility for new entrants to the field, but were challenged by less common structure type—both in terms of reduced performance for **A**/**E**/**P** predictions when training included **B**/**C**/**D** hosts, and in terms of these models' complete inability to make predictions for **B**/**C**/**D** host types. When investigating the results within the **A**/**E**/**P** set more deeply, two of the three *ortho*-substituted hosts (*i.e.* **AO1** and **AO3**) were among the worst predicted for both **SVM** and **RF**, further showing how these models may struggle to move beyond structure types well represented in the training set. We hope that the workflow described here will prove to be a helpful starting point for other systems that lack large public datasets.

Author contributions

F. H. and J. F. V. H. conceptualized, supervised, provided project administration and acquired funding. A. S. developed and investigated synthetic methodology. A. H. B. M. and D. R. T. developed software and investigated ML approaches. J. F. V. H. prepared the initial manuscript draft; all authors reviewed and edited the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

All python files necessary to perform the analysis described in this paper, as well as those necessary to generate the 3D voxelized representation of the calixarene hosts can be found on GitHub: github.com/JVH-YYC/Calixarenes.

All data supporting this work can be found in the manuscript and the associated supplementary information (SI). Supplementary information: experimental and computational methods, materials, details. LC-MS analysis of host library

synthesis and raw data for indicator displacement assays. See DOI: https://doi.org/10.1039/d6sc00479b.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada through Discovery Grants to JFVH (RGPIN-2019-06290) and FH (RGPIN-05118-2024) and by the University of Calgary Vice-President Research's office through a Catalyst Grant.

Notes and references

1 B. C. Haas, D. Kalyani and M. S. Sigman, Applying Statistical Modelling Strategies to Sparse Datasets in Synthetic Chemistry, *Sci. Adv.*, 2025, **11**, eadt3013.
2 H. Shalit Peleg and A. Ailo, Small Data Can Play a Big Role in Chemical Discovery, *Angew. Chem., Int. Ed.*, 2023, **62**, e202219070.
3 M. C. Ramos, C. J. Collison and A. D. White, A Review of Large Language Models and Autonomous Agents in Chemistry, *Chem. Sci.*, 2025, **16**, 2514.
4 L. M. Sigmund, M. Assante, M. J. Johansson, P.-O. Norrby, K. Jorner and M. Kabeshov, Computational Tools for the Prediction of Site- and Regioselectivity of Organic Reactions, *Chem. Sci.*, 2025, **16**, 5383.
5 P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman and C. W. Coley, Dataset Design for Building Models of Chemical Reactivity, *ACS Cent. Sci.*, 2023, **9**, 2196.
6 W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle and E. V. Anslyn, The Evolution of Data-Driven Modeling in Organic Chemistry, *ACS Cent. Sci.*, 2021, **7**, 1622.
7 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, The Protein Data Bank, *Nucleic Acids Res.*, 2000, **28**, 235.
8 B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum and A. R. Leach, The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods, *Nucleic Acids Res.*, 2024, **52**, D1180.
9 S. C. Smith, C. S. Horbaczewskyj, T. F. N. Tanner, J. J. Walder and I. J. S. Fairlamb, Automated Approaches, Reaction Parameterisation, and Data Science in Organometallic Chemistry and Catalysis: Towards Improving Synthetic Chemistry and Accelerating Mechanistic Understanding, *Digital Discovery*, 2024, **3**, 1467.
10 S. W. Krska, D. A. DiRocco, S. D. Dreher and M. Shevlin, The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis, *Acc. Chem. Res.*, 2017, **50**, 2976.
11 P. S. Kutchukian, J. F. Dropinski, K. D. Dykstra, B. Li, D. A. DiRocco, E. C. Streckfuss, L.-C. Campeau, T. Cernak, P. Vachal, I. W. Davies, S. W. Krska and S. D. Dreher, Chemistry Informer Libraries: A Chemoinformatics Enabled Approach to Evaluate and Advance Synthetic Methods, *Chem. Sci.*, 2016, **7**, 2604.
12 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, Bayesian Reaction Optimization as a Tool for Chemical Synthesis, *Nature*, 2021, **590**, 89.
13 S. Lin, S. Dikler, W. D. Blincoe, R. D. Ferguson, R. P. Sheridan, Z. Peng, D. V. Conway, K. Zawatzky, H. Wang, T. Cernak, I. W. Davies, D. A. DiRocco, H. Sheng, C. J. Welch and S. D. Dreher, Mapping the Dark Space of Chemical Reactions with Extended Nanomole Synthesis and MALDI-TOF MS, *Science*, 2018, **361**, eaar6236.
14 S. Pablo-García, Á. García, G. D. Akkoc, M. Sim, Y. Cao, M. Somers, C. Hattrick, N. Yoshikawa, D. Dworschak, H. Hao and A. Aspuru-Guzik, An Affordable Platform for Automated Synthesis and Electrochemical Characterization, *Device*, 2025, **3**, 100567.
15 F. Strieth-Kalthoff, H. Hao, V. Rathore, J. Derasp, T. Gaudin, N. H. Angello, M. Seifrid, E. Trushina, M. Guy, J. Liu, X. Tang, M. Mamada, W. Wang, T. Tsagaantsooj, C. Lavigne, R. Pollice, T. C. Wu, K. Hotta, L. Bodo, S. Li, M. Haddadnia, A. Wołos, R. Roszak, C. T. Ser, C. Bozal-Ginesta, R. J. Hickman, J. Vestfrid, A. Aguilar-Granda, E. L. Klimareva, R. C. Sigerson, W. Hou, D. Gahler, S. Lach, A. Warzybok, O. Borodin, S. Rohrbach, B. Sanchez-Lengeling, C. Adachi, B. A. Grzybowski, L. Cronin, J. E. Hein, M. D. Burke and A. Aspuru-Guzik, Delocalized, Asynchronous, Closed-Loop Discovery of Organic Laser Emitters, *Science*, 2025, **384**, eadk9227.
16 Work is beginning towards closing this gap. "SupraBank" (https://suprabank.org) has information regarding 2,145 compounds (as of October 14, 2025). For context, ChEMBL contains approximately 2.9 million.
17 M. A. Beatty, A. J. Selinger, Y. Li and F. Hof, Parallel Synthesis and Screening of Supramolecular Chemosensors That Achieve Fluorescent Turn-on Detection of Drugs in Saliva, *J. Am. Chem. Soc.*, 2019, **141**, 16763.
18 J. Chen, A. D. Gill, B. L. Hickey, Z. Gao, X. Cui, R. J. Hooley and W. Zhong, Machine Learning Aids Classification and Discrimination of Noncanonical DNA Folding Motifs by an Arrayed Host:Guest Sensing System, *J. Am. Chem. Soc.*, 2021, **143**, 12791.
19 J. M. Parrilla-Gutiérrez, J. M. Granda, J.-F. Ayme, M. D. Bajczyk, L. Wilbraham and L. Cronin, Electron Density-Based GPT for Optimization and Suggestion of Host–Guest Binders, *Nat. Comput. Sci.*, 2024, **4**, 200.
20 R. Reitalu, T. Jarg, R. Aav and M. Öeren, Predicting the Host–Guest Binding Gibbs Free Energy for Anion Guests, *ChemRxiv*, 2025, preprint, DOI: DOI: 10.26434/chemrxiv-2025-fp0ff.
21 M. H. Fitz-James and G. Cavalli, Molecular Mechanisms of Transgenerational Epigenetic Inheritance, *Nat. Rev. Genet.*, 2022, **23**, 325.
22 R. Liu, J. Wu, H. Guo, W. Yao, S. Li, Y. Lu, Y. Jia, X. Liang, J. Tang and H. Zhang, Post-Translational Modifications of

Histones: Mechanisms, Biological Functions, and Therapeutic Targets, *MedComm*, 2023, **4**, e292.
23 I. V. Bure, M. V. Nemtsova and E. B. Kuznetsova, Histone Modifications and Non-Coding RNAs: Mutual Epigenetic Regulation and Role in Pathogenesis, *Int. J. Mol. Sci.*, 2022, **23**, 5801.
24 B. D. Strahl and C. D. Allis, The Language of Covalent Histone Modifications, *Nature*, 2000, **403**, 41.
25 T. Jenuwein and C. D. Allis, Translating the Histone Code, *Science*, 2001, **293**, 1074.
26 C. S. Beshara, C. E. Jones, K. D. Daze, B. J. Lilgert and F. Hof, A Simple Calixarene Recognizes Post-Translationally Methylated Lysine, *ChemBioChem*, 2010, **11**, 63.
27 K. D. Daze, M. C. F. Ma, F. Pineux and F. Hof, Synthesis of New Trisulfonated Calix[4]Arenes Functionalized at the Upper Rim, and Their Complexation with the Trimethyllysine Epigenetic Mark, *Org. Lett.*, 2012, **14**, 1512.
28 K. D. Daze, T. Pinter, C. S. Beshara, A. Ibraheem, S. A. Minaker, M. C. F. Ma, R. J. M. Courtemanche, R. E. Campbell and F. Hof, Supramolecular Hosts That Recognize Methyllysines and Disrupt the Interaction between a Modified Histone Tail and Its Epigenetic Reader Protein, *Chem. Sci.*, 2012, **3**, 2695.
29 S. Tabet, S. F. Douglas, K. D. Daze, G. A. E. Garnett, K. J. H. Allen, E. M. M. Abrioux, T. T. H. Quon, J. E. Wulff and F. Hof, Synthetic Trimethyllysine Receptors That Bind Histone 3, Trimethyllysine 27 (H3K27me3) and Disrupt Its Interaction with the Epigenetic Reader Protein CBX7, *Bioorg. Med. Chem.*, 2013, **21**, 7004.
30 Y. Kimura, N. Saito, K. Hanada, J. Liu, T. Okabe, S. A. Kawashima, K. Yamatsugu and M. Kanai, Supramolecular Ligands for Histone Tails by Employing a Multivalent Display of Trisulfonated Calix[4]Arenes, *ChemBioChem*, 2015, **16**, 2599.
31 N. K. Pinkin, A. N. Power and M. L. Waters, Late Stage Modification of Receptors Identified from Dynamic Combinatorial Libraries, *Org. Biomol. Chem.*, 2015, **13**, 10939.
32 L. A. Ingerman, M. E. Cuellar and M. L. Waters, A Small Molecule Receptor That Selectively Recognizes Trimethyl Lysine in a Histone Peptide with Native Protein-like Affinity, *Chem. Commun.*, 2010, **46**, 1839.
33 R. Pinalli, G. Brancatelli, A. Pedrini, D. Menozzi, D. Hernández, P. Ballester, S. Geremia and E. Dalcanale, The Origin of Selectivity in the Complexation of N-Methyl Amino Acids by Tetraphosphonate Cavitands, *J. Am. Chem. Soc.*, 2016, **138**, 8569.
34 Y. Liu, L. Perez, M. Mettry, C. J. Easley, R. J. Hooley and W. Zhong, Self-Aggregating Deep Cavitand Acts as a Fluorescence Displacement Sensor for Lysine Methylation, *J. Am. Chem. Soc.*, 2016, **138**, 10746.
35 T. Hanauer, R. J. Hopkinson, K. Patel, Y. Li, D. Correddu, A. Kawamura, V. Sarojini, I. K. H. Leung and T. Gruber, Selective Recognition of the Di/Trimethylammonium Motif by an Artificial Carboxycalixarene Receptor, *Org. Biomol. Chem.*, 2017, **15**, 1100.
36 H. Peacock, C. C. Thinnes, A. Kawamura and A. D. Hamilton, Tetracyanoresorcin[4]Arene Selectively Recognises Trimethyllysine and Inhibits Its Enzyme-Catalysed Demethylation, *Supramol. Chem.*, 2016, **28**, 575.
37 I. N. Gober and M. L. Waters, Optimization of a Synthetic Receptor for Dimethyllysine Using a Biphenyl-2,6-Dicarboxylic Acid Scaffold: Insights into Selective Recognition of Hydrophilic Guests in Water, *Org. Biomol. Chem.*, 2017, **15**, 7789.
38 A. G. Mullins, N. K. Pinkin, J. A. Hardin and M. L. Waters, Achieving High Affinity and Selectivity for Asymmetric Dimethylarginine by Putting a Lid on a Box, *Angew. Chem., Int. Ed.*, 2019, **58**, 5282.
39 M. A. Gamal-Eldin and D. H. Macartney, Selective Molecular Recognition of Methylated Lysines and Arginines by Cucurbit[6]Uril and Cucurbit[7]Uril in Aqueous Solution, *Org. Biomol. Chem.*, 2013, **11**, 488.
40 M. Dionisio, G. Oliviero, D. Menozzi, S. Federici, R. M. Yebeutchou, F. P. Schmidtchen, E. Dalcanale and P. Bergese, Nanomechanical Recognition of N-Methylammonium Salts, *J. Am. Chem. Soc.*, 2012, **134**, 2392.
41 I. Alessandri, E. Biavardi, A. Gianoncelli, P. Bergese and E. Dalcanale, Cavitands Endow All-Dielectric Beads With Selectivity for Plasmon-Free Enhanced Raman Detection of Nε-Methylated Lysine, *ACS Appl. Mater. Interfaces*, 2016, **8**, 14944.
42 A. Shaurya, G. A. E. Garnett, M. J. Starke, M. C. Grasdal, C. C. Dewar, A. Y. Kliuchynskyi and F. Hof, An Easily Accessible, Lower Rim Substituted Calix[4]Arene Selectively Binds N,N-Dimethyllysine, *Org. Biomol. Chem.*, 2021, **19**, 4691.
43 M. Amezcua, J. Setiadi and D. L. Mobley, The SAMPL9 host–guest blind challenge: an overview of binding free energy predictive accuracy, *Phys. Chem. Chem. Phys.*, 2024, **26**, 9207.
44 T. Harren, T. Gutermuth, C. Grebner, G. Hessler and M. Rarey, Modern Machine-Learning for Binding Affinity Estimation of Protein–Ligand Complexes: Progress, Opportunities, and Challenges, *WIREs Comput. Mol. Sci.*, 2024, **14**, e1716.
45 D. D. Wang, W. Wu and R. Wang, Structure-Based, Deep-Learning Models for Protein-Ligand Binding Affinity Prediction, *J. Cheminf.*, 2024, **16**, 2.
46 A. Nigam, R. Pollice, M. F. D. Hurley, R. J. Hickman, M. Aldeghi, N. Yoshikawa, S. Chithrananda, V. A. Voelz and A. Aspuru-Guzik, Assigning Confidence to Molecular Property Prediction, *Expert Opin. Drug Discov.*, 2021, **16**, 1009.
47 D. Rogers and M. Hahn, Extended-Connectivity Fingerprints, *J. Chem. Inf. Model.*, 2010, **50**, 742.
48 T. R. Lane, D. H. Foil, E. Minerali, F. Urbina, M. K. Zorn and S. Ekins, Bioactivity Comparison across Multiple Machine Learning Algorithms Using over 5000 Datasets for Drug Discovery, *Mol. Pharm.*, 2021, **18**, 403.
49 RDKit: Open-source cheminformatics. https://www.rdkit.org.
50 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss,

V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825.

51 J. Jiménez, M. Škalič, G. Martínez-Rosell and G. De Fabritiis, KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks, *J. Chem. Inf. Model.*, 2018, **58**, 287.

52 M. M. Stepniewska-Dziubinska, P. Zielenkiewicz and P. Siedlecki, Development and Evaluation of a Deep Learning Model for Protein–Ligand Binding Affinity Prediction, *Bioinformatics*, 2018, **34**, 3666.

53 D. Kuzminykh, D. Polykovskiy, A. Kadurin, A. Zhebrak, I. Baskov, S. Nikolenko, R. Shayakhmetov and A. Zhavoronkov, 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks, *Mol. Pharm.*, 2018, **15**, 4378.

54 X. Huo, J. Xu, M. Xu and H. Chen, An Improved 3D Quantitative Structure-Activity Relationships (QSAR) of Molecules with CNN-Based Partial Least Squares Model, *Artif. Intell. Life Sci.*, 2023, **3**, 100065.

55 F. Imrie, T. E. Hadfield, A. R. Bradley and C. M. Deane, Deep Generative Design with 3D Pharmacophoric Constraints, *Chem. Sci.*, 2021, **12**, 14577.

56 D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, Convolutional Networks on Graphs for Learning Molecular Fingerprints, in *Advances in Neural Information Processing Systems*, ed C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, Curran Associates, Inc, 2015, Vol. 28.

57 Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang and M. Zheng, Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism, *J. Med. Chem.*, 2020, **63**, 8749.

58 L. G. Cannell, The Bromodesulfonation of Aromatic Sulfonate Salts. II. The Effect of Amino, Methoxy, Methyl and Nitro Substituents, *J. Am. Chem. Soc.*, 1957, **79**, 2932.

59 V. Yadav, R. Kumar, S. Srikrishna and V. Prasad, Desulfonylative Halogenation of Arene Sulfonyl Chlorides Using N-Halosuccinimides: Synthesis, Molecular Docking, and Anti-Alzheimer Activity, *Bioorg. Med. Chem. Lett.*, 2025, **128**, 130321.

60 J. G. Traynham, Aromatic Substitution Reactions: When You've Said Ortho, Meta, and Para You Haven't Said It All, *J. Chem. Educ.*, 1983, **60**, 937.

61 D.-S. Guo, V. D. Uzunova, X. Su, Y. Liu and W. M. Nau, Operational Calixarene-Based Fluorescent Sensing Systems for Choline and Acetylcholine and Their Application to Enzymatic Reactions, *Chem. Sci.*, 2011, **2**, 1722.

62 H. Li, S. Ilin, W. Wang, E. M. Duncan, J. Wysocka, C. D. Allis and D. J. Patel, Molecular Basis for Site-Specific Read-out of Histone H3K4me3 by the BPTF PHD Finger of NURF, *Nature*, 2006, **442**, 91.

63 P. V. Peña, F. Davrazou, X. Shi, K. L. Walter, V. V. Verkhusha, O. Gozani, R. Zhao and T. G. Kutateladze, Molecular Mechanism of Histone H3K4me3 Recognition by Plant Homeodomain of ING2, *Nature*, 2006, **442**, 100.

64 S. Iwase, B. Xiang, S. Ghosh, T. Ren, P. W. Lewis, J. C. Cochrane, C. D. Allis, D. J. Picketts, D. J. Patel, H. Li and Y. Shi, ATRX ADD Domain Links an Atypical Histone Methylation Recognition Mechanism to Human Mental-Retardation Syndrome, *Nat. Struct. Mol. Biol.*, 2011, **18**, 769.

65 R. M. Hughes, K. R. Wiggins, S. Khorasanizadeh and M. L. Waters, Recognition of Trimethyllysine by a Chromodomain Is Not Driven by the Hydrophobic Effect, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 11184.

66 X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith and A. E. Roitberg, TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials, *J. Chem. Inf. Model.*, 2020, **60**, 3408.

67 C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev and A. E. Roitberg, Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens, *J. Chem. Theory Comput.*, 2020, **16**, 4192.