Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2025, **16**, 10833

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

Data efficient molecular image representation learning using foundation models†

Yonatan Harnik, [a] Hadas Shalit Peleg,[a] Amit H. Bermano*[b] and Anat Milo *[a]

Deep learning (DL) in chemistry has seen significant progress, yet its applicability is limited by the scarcity of large, labeled datasets and the difficulty of extracting meaningful molecular features. Molecular representation learning (MRL) has emerged as a powerful approach to address these challenges by decoupling feature extraction and property prediction. In MRL, a deep learning network is first trained to learn molecular features from large, unlabeled datasets and then finetuned for property prediction on smaller specialized data. Whereas MRL methods have been widely applied across chemical applications, these models are typically trained from scratch. Herein, we propose that foundation models can serve as an advantageous starting point for developing MRL models. Foundation models are large models trained on diverse datasets capable of addressing various downstream tasks. For example, large language models like OpenAI's GPT-4 can be finetuned with minimal additional data for tasks considerably different from their training. Based on this premise we leveraged OpenAI's vision foundation model, CLIP, as the backbone for developing MoleCLIP, a molecular image representation learning framework. MoleCLIP requires significantly less molecular pretraining data to match the performance of state-of-the-art models on standard benchmarks. Furthermore, MoleCLIP outperformed existing models on homogeneous catalysis datasets, emphasizing its robustness to distribution shifts, which allows it to adapt effectively to varied tasks and datasets. This successful application of a general foundation model to chemical tasks highlights the potential of innovations in DL research to advance synthetic chemistry and, more broadly, any field where molecular property description is central to discovery.

Received 4th February 2025
Accepted 13th May 2025

DOI: 10.1039/d5sc00907c

rsc.li/chemical-science

Manuscript

The fast-paced field of deep learning (DL) provides new opportunities in chemical research due to the exquisite ability of DL frameworks to capture complex relationships.[1–3] Nevertheless, a significant barrier for expanding the application of DL in chemistry is the limited availability of large and reliable molecular datasets, which would ideally contain millions of datapoints.[3–5] This data availability issue is compounded by the requirement for labeled datasets, in which reliable chemical property values and reaction outcomes are associated with each molecule.[5,6] In certain unique cases these challenges can be resolved by producing large amounts of simulated data;[7,8] yet this is not relevant to most chemical applications where experimental data is required. Consequently, recent years have seen a push toward self-driving laboratories (SDL) to produce consistent and reliable data through high-throughput experimentation (HTE).[9–11] However, these platforms are not yet widely available, and the creation of millions of samples is still restrictive. The process of extracting key features that represent the data, known as featurization, is also restrictive because defining and extracting predictive molecular descriptors for a large dataset is resource intensive and often necessitates a high degree of experience with chemical systems.[12–14]

Molecular representation learning (MRL) seeks to decouple these data and labeling challenges to enable DL applications for chemistry.[15,16] In MRL, the featurization and prediction tasks are separated into two distinct stages using different datasets. The first stage, referred to as pretraining, focuses on training a deep encoder that serves to convert molecular data into general-purpose features.[3] The pretraining phase is typically performed on large datasets of unlabeled molecules by self-supervised learning. This pretrained encoder can then generate features for specific prediction tasks—a process also known as transfer learning.[17] Thus, MRL bypasses the need to use large, labeled datasets. Beyond its substantial efficiency in terms of time and resources, this approach leverages the ability of DL to identify hidden patterns, potentially reducing human bias.

In the MRL workflow, molecules are introduced to the model through molecular formats such as graphs, strings, or images.[15,16] Graphs, in which nodes represent atoms and edges

[a]*Department of Chemistry, Ben-Gurion University of the Negev, Beer Sheva, Israel. E-mail: anatmilo@bgu.ac.il*

[b]*School of Computer Science, Tel Aviv University, Tel Aviv, Israel. E-mail: amberman@tauex.ac.il*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5sc00907c

represent bonds, have garnered the most attention as an input representation for MRL due to their intuitive and compact depiction of molecules.[18–22] Among the graph-based MRL frameworks, the most prevalent are geometrical models in which the pretraining stage is focused on spatial properties such as bond lengths and bond angles.[18,19] String representations, such as SMILES and SELFIES, have also been broadly applied in DL for chemistry because they are lightweight, compact, and easy to handle.[23] MRL frameworks that rely on molecular string representations often employ transformers due to their efficiency as encoders.[24–27] Molecular images have been paid less attention compared to graphs and strings,[28,29] perhaps because images represent molecules by sparse matrices of pixels, making them less explicit and compact. However, a significant and overlooked advantage of developing molecular representation models with images as their input is the ability to leverage a vision foundation model as a powerful backbone.

Foundation models have emerged as a prominent field of AI in recent years.[30] These are models that were trained on large and varied datasets and can address various downstream tasks by transfer learning. The main factor that sets foundation models apart is their scale—their pretraining requires exceptionally large datasets and computational resources typically beyond the reach of most research groups. However, once foundation models are generated and openly shared, they can be finetuned at very low computational cost for different downstream tasks. Thus, we hypothesized that foundation models might serve as an advantageous starting point for MRL pretraining as part of a sequential workflow. Conceptually, this workflow would employ a general-purpose foundation model, which would undergo pretraining with molecular inputs, and would then be finetuned for specific chemistry-related tasks (Fig. 1a).

To the best of our knowledge, no existing methods have leveraged foundation models as the initial backbone for molecular encoders. A few studies have used foundation models indirectly, for example, pretrained large language models (LLMs) have been employed to couple molecular representations with information extracted from textual chemical descriptions.[31–33] We propose that building MRL models that are initialized from the weights of general-purpose foundation models, rather than trained from scratch on molecular data,

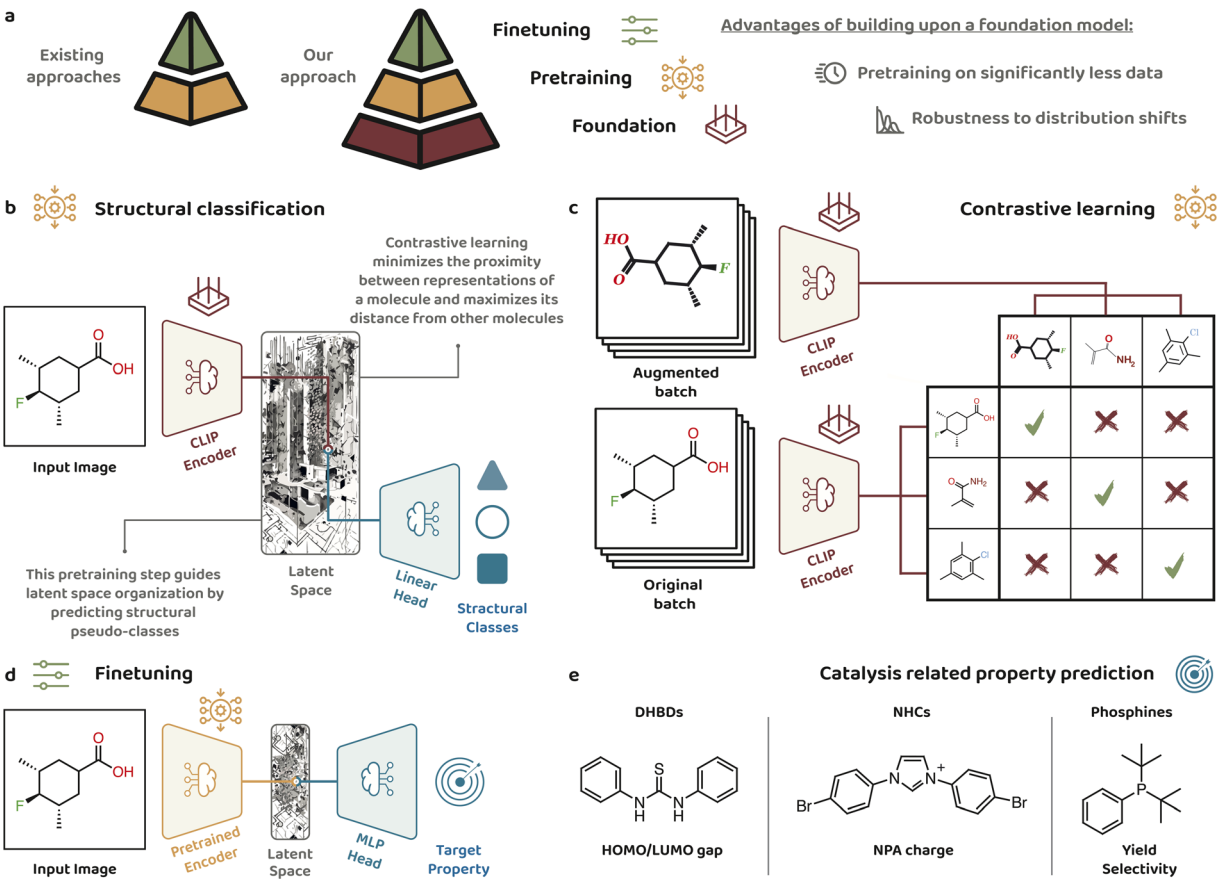

Fig. 1 **Overview of MoleCLIP conceptual framework, workflow, and data.** (a) The benefit of using foundation models as starting points for pretraining. (b and c) Illustration of the two pretraining tasks used in MoleCLIP. The weights of the MoleCLIP encoder are initialized from OpenAI CLIP as a foundation model and it is further pretrained by structural classification and contrastive learning. (d) Illustration of the finetuning process for property prediction. The encoder is initialized from the pretrained weights, and the decoder is a lightweight multi-perceptron (MLP) neural network trained from scratch for each task. (e) Three classes of catalysis-related molecules and their respective properties that were used for the evaluation of MoleCLIP's performance.

would lower the volume of molecular data required for pretraining. This strategy is very useful in a field such as chemistry, where data availability relies on experiments. More importantly, the use of foundation models can contribute to the model's robustness to distribution shifts, which is the ability to maintain performance on new tasks or domains that differ from those used in training.[30,34] This robustness arises from the extensive diversity of the foundation training data, which enables the model to capture a wide range of patterns.

Reliance on foundation models is feasible when their input is identical to the MRL encoder's input, and their architecture is similar. Applying this strategy to graph based MRL is currently impractical because general-purpose graph foundation models are still in their infancy.[35] It is also not trivial to use large language models (LLMs) as the foundation for string based MRL because the grammar of the chemical language, which could broadly be defined as chemical reactions, processes and properties,[36] is not inherently structured as a natural language. Based on these considerations, vision foundation models are currently an especially attractive starting point for MRL because image encoders do not require any modification to process images of molecules.

Herein, we test these ideas through the introduction of MoleCLIP, an MRL framework that accepts images of molecules as inputs and adopts a visual transformer architecture initialized with weights from OpenAI's CLIP (Contrastive Language-Image Pretraining) model, which was trained on a dataset of 400 million image and text pairs (Fig. 1b and c).[37] We found that MoleCLIP could be trained in a few-shot manner on significantly less molecular pretraining data compared to frameworks trained solely on molecules. MoleCLIP was evaluated on property prediction of MoleculeNet benchmarks[38] (Fig. 1d) and homogeneous catalysis datasets (Fig. 1e). It achieved comparable performance to state-of-the-art (SOTA) MRL models on MoleculeNet benchmarks despite being pretrained on considerably lower volumes of molecular data (see ESI Section S2c(i)†). On homogeneous catalysis datasets at the small data-size regime, MoleCLIP presented superior performance compared to ImageMol, which is currently the only reported image-based MRL model, and comparable to superior performance compared with GEM, a graph-based SOTA model. Moreover, we demonstrated the robustness to distribution shifts granted to MoleCLIP by relying on a vision foundation model.

Pretraining

The pretraining phase of MoleCLIP was performed on molecular image inputs generated by RDKit, which is a commonly used open-source package for cheminformatics.[39] The dataset selected for pretraining was ChEMBL-25, comprised of 1.9 M bioactive drug-like molecules.[40] During the molecular pretraining phase, the model was trained by two simultaneous tasks: structural classification (Fig. 1b) and contrastive learning (Fig. 1c). We selected these two tasks because they are complementary in addressing structural and image-related considerations for producing a molecular latent space, which is a vectorial space that captures the essential features of the input data.

The first task involved supervised classification of the molecules to structural classes, a task adopted from ImageMol,[29] a SOTA image-based MRL framework. The molecules were assigned structural fingerprints, which are low-computational cost, bit-vector molecular feature sets, and *K*-means clustering was performed across the dataset (see ESI Section S2a†). Then, pseudo-labels were assigned to each molecule based on their corresponding clusters. During this pretraining task, the model learned to classify each molecule to its structural pseudo-class (see Fig. 1b). The contrastive task followed SimCLR (Simple framework for Contrastive Learning of visual Representations), which entails creating augmented versions of each unlabeled image, for example by noise addition, rotation, or cropping.[41] Beyond these classical augmentation methods, generation-level augmentations could also be added to MoleCLIP by changing font types, font sizes, and line widths when generating the images by RDKit (for details see ESI Section S2b†). Both the original and augmented images were introduced to the model, which was trained to minimize the proximity in the latent space between pairs of images of the same molecule and to maximize the distance between images of different molecules (see Fig. 1c).

Finetuning

MoleculeNet benchmarks have been extensively used for molecular property prediction and drug discovery and are considered as standard for MRL evaluation; however, they exhibit certain biases.[42] These benchmarks consist of a diverse array of molecules containing various structural motifs; thus, are not well suited for evaluating model performance on narrow structural domains. Moreover, Deng *et al.* have highlighted the issue of excessive focus on assessing model performance by scaffold splitting in MoleculeNet benchmarks, where the molecular scaffolds that are most abundant are placed in the training set, and the less common scaffold motifs are placed in the validation and test sets.[42] Whereas this method ensures generalizability across different structural groups, it creates a bias towards models that perform well on inter-structural predictions. In practice, structurally similar molecules can exhibit significantly different potency or reactivity, a phenomenon known as an activity cliff, which poses a major challenge in molecular property prediction.[43] Homogeneous catalysis datasets are particularly prone to contain molecules within narrow structural domains and are often limited to tens of samples.[44] Based on these considerations, in addition to evaluating MoleCLIP on four MoleculeNet benchmarks, we tested it across four catalysis-related datasets of varying sizes containing molecules with either experimental or computational labels (Fig. 1e).

As the first catalysis case study, we chose a dataset of dual-hydrogen bond donors (DHBDs), a class of organocatalysts known for their effectiveness in a range of enantioselective reactions.[45] We selected 6994 combinatorically enumerated DHBD molecules from OSCAR, a comprehensive repository of organocatalysts by the Corminboeuf group.[46] The target property for finetuning by MoleCLIP was the density functional

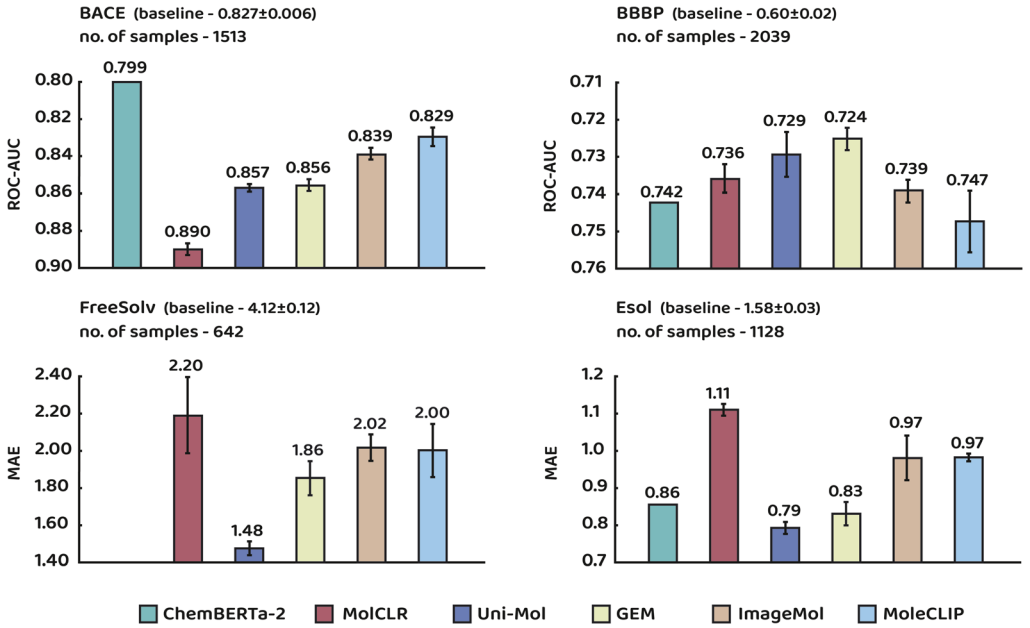

Fig. 2 **Comparison to state-of-the-art (SOTA).** Performance of MoleCLIP on MoleculeNet benchmarks compared to SOTA models. Baseline evaluations were performed using fixed representations (extended connectivity fingerprints). Evaluations were conducted using scaffold splitting, with error bars representing standard deviation based on three repeats. Despite lower pretraining data volume (presented in Table S4†), MoleCLIP achieves comparable performance to the reported performance metrics of SOTA models.

theory (DFT) computed gap between the highest occupied and lowest unoccupied molecular orbitals (HOMO/LUMO gap) extracted from OSCAR.

Another category of organocatalysts selected for evaluation was N-heterocyclic carbenes (NHCs). We curated a dataset of 95 NHC catalysts that were used in experimental settings.[47] The model was finetuned to predict DFT-calculated natural population analysis (NPA) charges (further details on the data collection and DFT calculations are available in the ESI Section S3a(iii)†). We chose to model NPA charges across the C–H bond of NHC pre-catalysts and across the C–C bond of a reactive intermediate common to numerous NHC-catalyzed reactions known as the Breslow intermediate.[48]

The last class of molecules we examined were organophosphines, which are widely used in transition-metal-catalyzed reactions as ligands to control reactivity and selectivity. For this class, two experimental datasets were selected. The first was produced by the Doyle group using HTE and consisted of 90 phosphine ligands and the target property was the yield each phosphine affords across five Ni-catalyzed Suzuki reactions (ESI Section S3a(iv)†).[49] A second dataset produced by the Sigman group was focused on the prediction of enantioselectivity across 37 different phosphines used as ligands for a Pd-catalyzed Suzuki reaction (ESI Section S3a(v)†).[50]

We compared the performance of MoleCLIP with established MRL frameworks and fingerprints-based baseline evaluation. Evaluating models against fixed representations such as fingerprints provides a solid baseline for assessing MRL capabilities.[42] Unlike learned representations, which are dynamically optimized during finetuning to capture task-relevant molecular features, fixed representations are predefined and are not adapted to specific downstream tasks. On MoleculeNet benchmarks, MoleCLIP achieved comparable performance to several SOTA MRL models despite being pretrained on significantly less molecular data (see Fig. 2 and ESI Table S4†). The ability of MoleCLIP to achieve competitive performance across the benchmarks, even with limited training data, can be attributed to its initialization with CLIP as a foundation model. We presume this stratified workflow—from CLIP weights to pretraining on molecular images, and finally to finetuning for downstream tasks—enabled MoleCLIP to effectively leverage a general-purpose image model for chemical applications.

On the catalysis datasets, MoleCLIP was compared to two representative MRL models: ImageMol,[29] an image-based framework, and GEM,[19] a graph-based framework. For reference, a baseline using a fixed fingerprint representation was also included (see also Table S5†). Despite being pretrained on significantly smaller volumes of molecular data (Fig. 3a), MoleCLIP demonstrated superior performance in most cases (Fig. 3b). Focusing on the three smaller datasets—NHCs, phosphines-yield, and phosphines-selectivity—each consisting of only a few dozen samples, MoleCLIP consistently outperformed ImageMol. It also achieved comparable or better performance than GEM, with a statistically significant improvement observed on the phosphines-selectivity dataset (Fig. 3b). Notably, MoleCLIP achieved this level of performance compared to GEM, even though GEM benefits from explicit geometric information that is not available to MoleCLIP.

We attribute MoleCLIP's improved performance to its ability to generalize effectively when transitioning from a diverse array of molecules in the pretraining stage to a narrower domain of structurally and mechanistically similar molecules in finetuning.

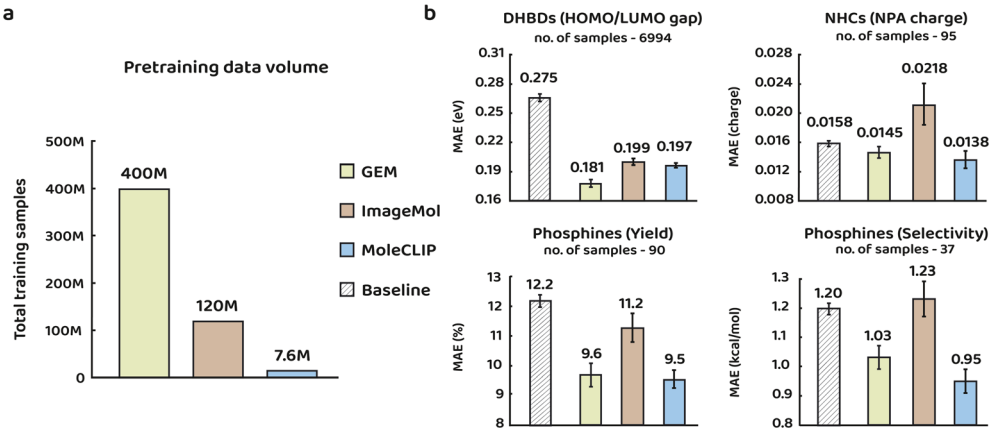

Fig. 3 **Catalysis-related datasets.** (a) The total number of pretraining samples is calculated as the product of the number of epochs and the number of molecules in the dataset. This comparison highlights the differences in the amount of data seen during pretraining across MoleCLIP (1.9 million samples, 4 epochs), ImageMol (10 million samples, 12 epochs), and GEM (20 million samples, 20 epochs). (b) Performance comparison of MoleCLIP, ImageMol, and GEM on the catalysis datasets. Baseline evaluations were performed using fixed representations (extended connectivity fingerprints). A mean absolute error (MAE) metric was used for all the datasets. Evaluations were conducted using random splitting, with error bars representing a 95% confidence interval. MoleCLIP consistently outperforms ImageMol on all datasets and surpasses GEM on three out of four. We note that the performance difference between MoleCLIP and GEM on the phosphine-yield and NHCs dataset is not statistically significant.

MoleCLIP's capacity to perform better on downstream tasks in domains that differ from its original training could be defined as robustness against distribution shifts, which we attribute to its reliance on a foundation model as a backbone.

Domain-focused pretraining

Building upon the success of continued pretraining, we hypothesized that MoleCLIP's performance might be further improved by adding a domain-focused pretraining step. Buried volume values (Fig. 4a) have been found as strong predictors of phosphine activity cliffs;[49] therefore, we speculated that a sequential strategy with an additional pretraining step focused on predicting buried volumes could improve model accuracy. This pretraining step was performed on a dataset of 1540 molecules along with their respective DFT-calculated buried volumes from the Kraken database curated by the Sigman and Aspuru-Guzik groups (see ESI Section S4b† for further details).[51]

The resulting model, which we refer to as $\text{MoleCLIP}_{\text{BV}}$, was evaluated against the phosphine yield and selectivity datasets. In both cases, $\text{MoleCLIP}_{\text{BV}}$ outperformed MoleCLIP's primary model (see Fig. 4a). We note that even in the absence of this additional continued training step, MoleCLIP not only exceeded ImageMol and GEM (see Fig. 3b), but its mean absolute error (MAE) was below the 1 kcal mol^{-1} chemical accuracy limit, which is considered the standard for realistic chemical predictions.[52] Nevertheless, this example emphasized the capacity of domain-related knowledge to significantly improve prediction accuracy (from an MAE of 0.95 to 0.82 kcal mol^{-1}, see Fig. 4a). Moreover, it illustrates the power of deploying a well-designed stratified workflow, going from a very general model, and gradually adding pretraining steps on smaller datasets with more accurate labels and an increasing relevance to a specific target. This outcome represents a promising future direction for refining pretraining workflows to better align with prediction targets in a broad range of domains.

Robustness to distribution shifts

To support our claim that MoleCLIP is more robust to distribution shifts, we set out to evaluate how it would handle molecular images that represent the data in a manner that was not introduced during pretraining. The letter R is often used in images of molecules as a general surrogate for functional groups, which are common atomic or molecular motifs; however, these groups are explicitly provided when training ML models. Therefore, replacing functional groups by the letter R in the input images could serve as an exquisite example of a distribution shift during finetuning on the target datasets. Namely, the model has seen images of molecules during pretraining but has never encountered an R as a substituent.

For this analysis, we selected the two regression MoleculeNet benchmarks, FreeSolv and Esol. We replaced functional groups in the datasets with numbered R-groups (Fig. 4b) and performed model finetuning using MoleCLIP and ImageMol (see ESI Section S4a† for further details). As expected, the accuracy of both models declined on the R-modified datasets (Table S6†). However, a clear trend emerged when comparing the relative performance of MoleCLIP and ImageMol, with MoleCLIP exhibiting a significantly smaller accuracy drop. This highlights MoleCLIP's resilience to new image types, supporting our claim that its robust foundation enables it to effectively handle distribution shifts.

Model limitations

The use of image-based representations in the MoleCLIP framework poses several inherent limitations. First, images are

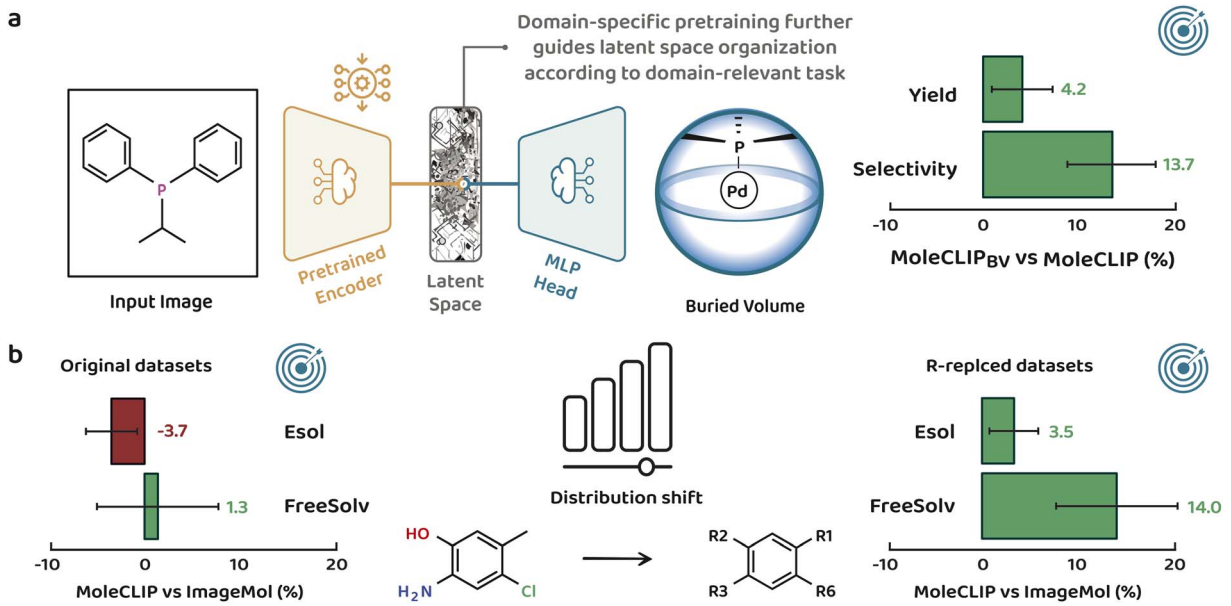

Fig. 4 **Added value of stratified pretraining.** (a) Illustration of the domain-specific continued pretraining methodology: the prediction of DFT-calculated buried volume values for a set of phosphine molecules was added to MoleCLIP as a pretraining stage. The domain-focused model, MoleCLIP$_{BV}$, showed statistically significant superior performances on phosphine yield and selectivity prediction dataset compared to the MoleCLIP primary model. The improvement was from an MAE of 9.5% to 9.1% in yield (a relative gap of 4.2%), and from an MAE of 0.95 to 0.82 kcal mol^{-1} in $\Delta\Delta G^{\ddagger}$ (a relative gap of 13.7%). The error bars represent a 95% confidence interval. (b) Analysis of MoleCLIP's robustness to distribution shifts: MoleCLIP and ImageMol were finetuned on original and R-replaced Esol and FreeSolv datasets. The relative gap in performance between MoleCLIP and ImageMol is larger in the R-replaced datasets, indicating MoleCLIP's better robustness to distribution shifts. Evaluations were conducted using random splitting, with error bars representing a 95% confidence interval.

less flexible than graph-based representations. Whereas graphs can easily embed additional spatial or chemical knowledge as atom or bond attributes, images lack a straightforward mechanism for encoding such details. Another limitation is a possible performance drop for very large molecules; however, we note that large molecules are generally more challenging to model across all representation types. For image representations, certain structural regions of very large molecules may appear blurred or visually ambiguous, which can hinder accurate interpretation of the full molecular structure (see ESI Section S4d(ii)†). This issue is unlikely to affect most of the practical chemical space yet remains a relevant concern. Another limitation of image-based workflows is that, compared to more compact formats such as SMILES strings or molecular graphs, image files are usually heavier, presenting challenges in terms of storage, memory usage, and training efficiency. Despite these limitations, the ability to leverage powerful foundation models, together with the prevalence of molecular images in the chemical literature, makes image-based representations a highly practical and promising approach for molecular representation learning.

Prior biases in the face of new evidence[53]

In this study we developed a powerful molecular image representation learning model using OpenAI's CLIP as a backbone. Our findings clearly demonstrate the efficiency of implementing a stratified learning strategy, which starts with a broad image foundation model, followed by few-shot continuous pretraining steps on molecular images toward specific downstream targets. This strategy not only enables high accuracy with minimal data volumes and computational cost, but it also provides advantages in handling tasks and data that are different from those used in pretraining. One might view MoleCLIP's flexibility toward shifts in data and tasks as reminiscent of children's ability to assimilate new information and update their beliefs. Compared to adults, young children are able to learn a wider variety of linguistic distinctions,[54] are better at distinguishing between faces of non-human primates,[55] excel at imagining new uses for tools,[56] and are more prone to infer initially unlikely causal hypotheses from a pattern of new evidence.[53] In this vein, we speculate that the large foundation model with which MoleCLIP was initialized can provide an extensive semantic latent space for embedding new images. Likewise, the shorter learning phase on a smaller set of molecular images does not entrench biases with respect to new downstream tasks and data compared to models trained from scratch solely on a vast set of molecules. Ultimately, we hope that the disclosed gains of building upon an image foundation model will illustrate the broad implications of general-purpose foundation models as starting points for DL in chemistry.

Methods

Architecture

MoleCLIP relies on OpenAI's CLIP,[37] a visual transformer (ViT) based foundation model that offers a solid starting point for

MoleCLIP's continued pretraining phase on molecular images. Specifically, the ViT-B/16 variant of CLIP was used, which includes 12 transformer encoder layers and processes images with dimensions of 224 × 224 pixels, utilizing a patch size of 16 × 16 pixels.

Pretraining data

The pretraining of MoleCLIP was performed on molecular image inputs generated by converting SMILES[57] to 224 × 224 images using RDKit.[39] ChEMBL-25, comprised of 1 870 421 bioactive drug-like molecules, was used as an unlabeled dataset for pretraining (see Data availability in ESI†).[58] To enhance the model's ability to embed molecules effectively, two distinct pretraining tasks were combined during the pretraining phase: supervised structural classification and self-supervised contrastive learning. This combination was intended to enable the model to capture distinctions at the scaffold level, as well as finer molecular details. To evaluate the individual contribution of each task toward input data encoding, we provide an analysis and visualization of the resulting embedding space (see ESI Section 4d(i)†).

Structural classification

Based on a task developed for ImageMol,[29] we incorporated a structural classification pretraining task to teach the encoder to differentiate between various structural groups within the embedding space. We extracted 166-bit MACCS (Molecular ACCess System) fingerprints[59] from SMILES using RDKit for each of the molecules in the dataset.[39] We then employed the K-means algorithm to cluster the molecules using the fingerprints as features; thus, assigning pseudo structural classes to the molecules in the dataset. To determine the optimal number of clusters (K), we tested K values ranging from 3 to 3000, where the optimal K was identified using the knee-point detection algorithm.[60] The knee-point curve for the ChEMBL-25 dataset, shown in Fig. S2,† indicated an optimal K value of 300. Inspired by ImageMol, we set clustering labels for our primary model at $K = 300$ and $K = 3000$, aiming to capture both coarse and fine-grained patterns in the data. As an ablation study, we tested also a case with clustering labels $K = 30$ and $K = 300$ (see ESI Section 4c(ii)†).

Each molecular image was assigned a structural class pseudo-label according to the cluster to which it belonged. As depicted in Fig. 1b, a linear head was added atop the embedding layer to predict the class. The training involved calculating cross-entropy loss between the model's predictions and the assigned pseudo-labels. The structural classification loss for a batch of samples is given by:

$$\mathcal{L}_{\mathrm{SC}} = -\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\,\log\frac{\exp(x_{n,c})}{\sum_{i=1}^{C}\exp(x_{n,i})}$$

where N is the batch size, C is the number of classes, x represents the predicted values, and y is the pseudo label.

Contrastive learning

Inspired by SimCLR,[41] pairs of original and augmented molecular images were generated using RDKit. The model then encoded both sets, training to minimize the embedding distance between images of the same molecule while maximizing the distance between images of different molecules. In addition to classical augmentation methods such as rotation, blurring, noise addition, the use of RDKit for image generation allowed creating of generation-level augmentations, such as font type, line width, and font size, as detailed in ESI Section S2b(i).†

During training, the batches were designed to include molecules from different and similar structural classes. This approach helped the model distinguish between different and alike structures. Batches of 32 samples from the same class were initially created and then randomly combined to form final training batches of 256 samples. The contrastive loss was computed similarly to the method used in CLIP,[37] where the contrastive loss function for a batch of samples (adapted from Zhai *et al.*) is defined as:[61]

$$\mathcal{L}_{\mathrm{CL}} = -\frac{1}{2N}\sum_{n=1}^{N}\left(\log\frac{\exp\left(\frac{1}{\tau}I_n^{\mathrm{o}}\cdot I_n^{\mathrm{a}}\right)}{\sum_{i=1}^{N}\exp\left(\frac{1}{\tau}I_n^{\mathrm{o}}\cdot I_i^{\mathrm{a}}\right)} + \log\frac{\exp\left(\frac{1}{\tau}I_n^{\mathrm{o}}\cdot I_n^{\mathrm{a}}\right)}{\sum_{i=1}^{N}\exp\left(\frac{1}{\tau}I_i^{\mathrm{o}}\cdot I_n^{\mathrm{a}}\right)}\right)$$

where N is the batch size, I^{o} is the normalized embedding of an original image, I^{a} is the normalized embedding of an augmented image, and $\frac{1}{\tau}$ is the temperature scaling factor.

Pretraining process

The encoder was pretrained using four Nvidia-T4 GPUs (64 GB RAM) over four epochs on the ChEMBL-25 dataset. The training was performed using the Adam optimizer, with a learning rate of 5×10^{-6} for the encoder (100 times lower than the CLIP learning rate to prevent catastrophic forgetting) and a rate of 0.01 for the structural classification linear head. We applied a weight decay of 0.1 and a batch size of 256. The total loss for pretraining is defined as the summation of the structural classification (SC) and contrastive learning (CL) losses:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SC}} + \mathcal{L}_{\mathrm{CL}}$$

Finetuning

Finetuning was conducted on datasets of property-labeled molecules, with performance evaluated based on the models' property prediction capabilities. A 3-layer, 512-dimensional multilayer perceptron (MLP) was added on top of the pretrained encoder. The encoder and the MLP head were trained simultaneously using different optimized learning rates. MoleCLIP employed the Adam optimizer throughout all training sessions, with the following parameters: $\beta 1 = 0.9$, $\beta 2 = 0.98$, $\varepsilon = 1 \times 10^{-6}$, weight decay $= 1 \times 10^{-5}$.

For each finetuning task, hyperparameter optimization was performed by exploring a range of learning rates and image

augmentation intensities. A similar approach was used for finetuning and evaluating GEM and ImageMol, utilizing pretrained weights and hyperparameter optimization as provided by the respective authors. Further details on the datasets used for finetuning, training procedures, augmentations, splitting, and evaluation protocols are available in ESI Section S3.†

Domain-focused pretraining

The domain-focused pretraining session was performed using the Kraken dataset, which includes 1540 literature-sourced molecules and their corresponding DFT-calculated properties.[51] This continued pretraining phase was initialized from the ChEMBL-pretrained weights of the MoleCLIP primary model. It was executed similarly to a finetuning process, where the MoleCLIP encoder and an additional 3-layer MLP prediction head (512-dimensional) were trained simultaneously. The focused pretraining was run for 300 epochs with a constant learning rate of 5×10^{-6}, weight decay of 1×10^{-5}, and batch size of 64. Further details on the domain-focused pretraining process and the following finetuning process are provided in ESI Section S4b.†

Data availability

Code, datasets and results are available at **https://github.com/Milo-group/MoleCLIP**. Pretrained model weights are available at **https://zenodo.org/records/13826016**. See ESI Data availability section† for further details.

Author contributions

The idea for this work was developed by Y. H., A. H. B. and A. M.; the model was designed by Y. H. and A. H. B. with input from A. M.; coding and training were performed by Y. H.; the datasets were processed by Y. H. and H. S. P.; the computations for the NHC dataset and their curation were perfomed by H. S. P.; Y. H. and A. M. wrote the manuscript with input from A. H. B. and H. S. P.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

A. M. wishes to thank the Israel Science Foundation (ISF) for their generous support (grant no. 2252/21) and Y. H. wishes to thank the Kreitman School of Advanced Graduate Studies for the Chemotech fellowship. The authors thank the Israel Data Science and AI Initiative (IDSAI) for their generous support through cloud computation resources.

References

1 H. Shalit Peleg and A. Milo, *Angew. Chem., Int. Ed.*, 2023, **62**, e202219070.
2 B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, Y. Zhu, J. Liu, B. Zhang and G. W. Wei, *Chem. Rev.*, 2023, **123**, 8736–8780.
3 A. C. Mater and M. L. Coote, *J. Chem. Inf. Model.*, 2019, **59**, 2545–2559.
4 W. P. Walters and R. Barzilay, *Acc. Chem. Res.*, 2021, **54**, 263–270.
5 L. Y. Chen and Y. P. Li, *Beilstein J. Org. Chem.*, 2024, **20**, 2476–2492.
6 M. L. Schrader, F. R. Schäfer, F. Schäfers and F. Glorius, *Nat. Chem.*, 2024, **16**(4), 491–498.
7 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
8 J. Burés and I. Larrosa, *Nature*, 2023, **613**, 689–695.
9 M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik and J. E. Hein, *Commun. Chem.*, 2021, **4**(1), 1–12.
10 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, *Nature*, 2023, **624**(7992), 570–578.
11 E. S. Isbrandt, R. J. Sullivan and S. G. Newman, *Angew. Chem., Int. Ed.*, 2019, **58**, 7180–7191.
12 E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chem. Soc. Rev.*, 2020, **49**, 3525–3564.
13 S. Singh and R. B. Sunoj, *Acc. Chem. Res.*, 2023, **56**, 402–412.
14 D. S. Wigh, J. M. Goodman and A. A. Lapkin, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1603.
15 Y. Harnik and A. Milo, *Chem. Sci.*, 2024, **15**, 5052–5055.
16 Z. Li, M. Jiang, S. Wang and S. Zhang, *Drug Discovery Today*, 2022, **27**, 103373.
17 X. Li and D. Fourches, *J. Cheminf.*, 2020, **12**, 1–15.
18 G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang and G. Ke, *ChemRxiv*, 2022, preprint, DOI: **10.26434/chemrxiv-2022-jjm0j-v4**.
19 X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu and H. Wang, *Nat. Mach. Intell.*, 2022, **4**, 127–134.
20 Y. Wang, J. Wang, Z. Cao and A. Barati Farimani, *Nat. Mach. Intell.*, 2022, **4**(3), 279–287.
21 Y. Rong, Y. Bian, T. Xu, W. Xie, Y. WEI, W. Huang and J. Huang, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 12559–12571.
22 K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen and R. Barzilay, *J. Chem. Inf. Model.*, 2019, **59**, 3370–3388.
23 M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka, R. F. Lameiro, D. Lemm, A. Lo, S. M. Moosavi, J. M. Nápoles-Duarte, A. K. Nigam, R. Pollice, K. Rajan, U. Schatzschneider, P. Schwaller, M. Skreta, B. Smit, F. Strieth-Kalthoff, C. Sun, G. Tom, G. Falk von Rudorff, A. Wang, A. D. White, A. Young, R. Yu and A. Aspuru-Guzik, *Patterns*, 2022, **3**, 100588.
24 A. Yüksel, E. Ulusoy, A. Ünlü and T. Doğan, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 025035.
25 S. Chithrananda, G. Grand and B. R. Deepchem, *arXiv*, 2020, preprint, arXiv:2010.09885, DOI: **10.48550/arXiv.2010.09885**.

26 S. Wang, Y. Guo, Y. Wang, H. Sun and J. Huang, *ACM-BCB 2019 – Proc Int Conf Bioinform, Computational Biology and Health Informatics*, 2019, pp. 429–436.
27 W. Ahmad, E. Simon, S. Chithrananda, G. Grand and B. Ramsundar, *arXiv*, 2022, preprint, arXiv:2209.01712, DOI: 10.48550/arXiv.2209.01712.
28 Y. Li, B. Liu, J. Deng, Y. Guo and H. Du, *Briefings Bioinf.*, 2024, **25**(4), bbae294.
29 X. Zeng, H. Xiang, L. Yu, J. Wang, K. Li, R. Nussinov and F. Cheng, *Nat. Mach. Intell.*, 2022, **4**, 1004–1016.
30 R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou and P. Liang, *arXiv*, 2021, preprint, arXiv:2108.07258, DOI: 10.48550/arXiv.2108.07258.
31 J. J. Ben-Joseph and T. Oates, Cold Spring Harbor Laboratory, *bioRxiv*, 2023, preprint, DOI: 10.1101/2023.11.11.566721.
32 S. Balaji, R. Magar, Y. Jadhav and A. B. Farimani, *arXiv*, 2023, preprint, arXiv:2310.03030, DOI: 10.48550/arXiv.2310.03030.
33 Y. Deng, S. S. Ericksen and A. Gitter, *arXiv*, 2024, preprint, arXiv:2410.20182, DOI: 10.48550/arXiv.2108.07258.
34 C. J. Reed, X. Yue, A. Nrusimha, S. Ebrahimi, V. Vijaykumar, R. Mao, B. Li, S. Zhang, D. Guillory, S. Metzger, K. Keutzer and T. Darrell, *Proc IEEE Int Conf Comput Vis*, WACV, 2022, pp. 1050–1060.
35 J. Liu, C. Yang, Z. Lu, J. Chen, Y. Li, M. Zhang, T. Bai, Y. Fang, L. Sun, P. S. Yu and C. Shi, *arXiv*, 2023, preprint, arXiv:2310.11829, DOI: 10.48550/arXiv.2310.11829.
36 P. Schwaller, B. Hoover, J. L. Reymond, H. Strobelt and T. Laino, *Sci. Adv.*, 2021, **7**, eabe4166.
37 A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, *Proc. Mach. Learn. Res.*, 2021, vol. 139, pp. 8748–8763.
38 Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, *Chem. Sci.*, 2018, **9**, 513–530.
39 G. Landrum, P. Tosco, B. Kelley, R. Rodriguez-Schmidt, D. Cosgrove, S. Riniker, P. Gedeck, R. Vianello, N. Schneider, E. Kawashima, N. Dan, G. Jones, A. Dalke, B. Cole, M. Swain, S. Turk, A. Savelyev, A. Vaucher, M. Wójcikowski, I. Take, D. Probst, K. Ujihara, V. F. Scalfani, G. Godin, J. Lehtivarjo, A. Pahl, R. Walker, F. Berenger, J. D. Biggs and A. Stretton, *Zenodo repository*, 2023, DOI: 10.5281/zenodo.8413907.
40 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.
41 T. Chen, S. Kornblith, M. Norouzi and G. Hinton, in *Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
42 J. Deng, Z. Yang, H. Wang, I. Ojima, D. Samaras and F. Wang, *Nat. Commun.*, 2023, **14**(1), 1–20.
43 D. Van Tilborg, A. Alenicheva and F. Grisoni, *J. Chem. Inf. Model.*, 2022, **62**, 5938–5951.
44 P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman and C. W. Coley, *ACS Cent. Sci.*, 2023, **9**, 2196–2204.
45 J. Werth and M. S. Sigman, *J. Am. Chem. Soc.*, 2020, **142**, 16382.
46 S. Gallarati, P. van Gerwen, R. Laplaza, S. Vela, A. Fabrizio and C. Corminboeuf, *Chem. Sci.*, 2022, **13**, 13782–13794.
47 D. M. Flanigan, F. Romanov-Michailidis, N. A. White and T. Rovis, *Chem. Rev.*, 2015, **115**, 9307–9387.
48 S. C. Gadekar, V. Dhayalan, A. Nandi, I. L. Zak, M. S. Mizrachi, S. Kozuch and A. Milo, *ACS Catal.*, 2021, **11**, 14561–14569.
49 S. H. Newman-Stonebraker, S. R. Smith, E. Borowski, E. Peters, T. Gensch, H. C. Johnson, M. S. Sigman and A. G. Doyle, *Science*, 2021, **374**, 301–308.
50 Z. L. Niemeyer, A. Milo, D. P. Hickey and M. S. Sigman, *Nat. Chem.*, 2016, **8**(6), 610–617.
51 T. Gensch, G. Dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman and A. Aspuru-Guzik, *J. Am. Chem. Soc.*, 2022, **144**, 1205–1217.
52 M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K. R. Müller and K. Burke, *Nat. Commun.*, 2020, **11**(1), 1–11.
53 A. Gopnik, S. O'Grady, C. G. Lucas, T. L. Griffiths, A. Wente, S. Bridgers, R. Aboody, H. Fung and R. E. Dahl, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 7892–7899.
54 P. K. Kuhl, *Nat. Rev. Neurosci.*, 2004, **5**(11), 831–843.
55 O. Pascalis, M. De Haan and C. A. Nelson, *Science*, 2002, **296**, 1321–1323.
56 T. P. German and M. A. Defeyter, *Psychonomic Bull. Rev.*, 2000, **7**, 707–712.
57 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
58 CHEMBL database release 25, 2019.
59 J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273–1280.
60 V. Satopää, J. Albrecht, D. Irwin and B. Raghavan, in *Proc. – Int. Conf. Distrib. Comput. Syst.*, 2011, pp. 166–171.
61 X. Zhai, B. Mustafa, A. Kolesnikov and L. Beyer, in *Proc. IEEE Int. Conf. Comput. Vis.*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 11941–11952.