Pre-trained molecular representations enable antimicrobial discovery

Roberto Olayo-Alarcon[1,2], Martin K. Amstalden[3], Annamaria Zannoni[6], Medina Bajramovic[1], Cynthia M. Sharma[6], Ana Rita Brochado[3,4,5], Mina Rezaei[1], and Christian L. Müller[1,2,7]

[1]Department of Statistics, Ludwig-Maximilians-Universität München
[2]Institute of Computational Biology, Helmholtz Munich
[3]Department of Microbiology, Biocenter, Julius-Maximilians-Universität Würzburg
[4]Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), University of Tübingen
[5]Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen
[6]Department of Molecular Infection Biology II, Institute of Molecular Infection Biology (IMIB), Julius-Maximilians-Universität Würzburg
[7]Center for Computational Mathematics, Flatiron Institute, New York

Abstract

The rise in antimicrobial resistance poses a worldwide threat, reducing the efficacy of common antibiotics. Yet, determining the antimicrobial activity of new chemical compounds through experimental methods is still a time-consuming and costly endeavor. Compound-centric deep learning models hold the promise to speed up the search and prioritization process. Here, we introduce a lightweight computational strategy for antimicrobial discovery that builds on `MolE` (`M`olecular representation through redundancy reduced `E`mbedding). `MolE` is a non-contrastive self-supervised Graph Neural Network framework that leverages unlabeled chemical structures to learn task-independent molecular representations. By combining a pre-trained `MolE` representation with experimentally validated compound-bacteria activity data, we build an antimicrobial prediction model that re-discovers recently reported growth-inhibitory compounds that are structurally distinct from current antibiotics. Using the model as a compound prioritization strategy, we identify and experimentally confirm three human-targeted drugs as growth-inhibitors of Staphylococcus aureus, highlighting `MolE`'s potential to accelerate the discovery of new antibiotics.

Introduction

The development of novel antibiotics is a priority given the widespread dissemination of pathogenic strains that are resistant to current treatments [27]. Currently, novel therapeutic candidates are often first identified by screening large chemical libraries. However, the success of these screenings is often limited, with a typical hit rate between 1-3% being a common occurrence [35, 37]. This issue is compounded by the high cost of the experiments owing to the large size of the chemical libraries being evaluated (from thousands to millions of molecules). Furthermore, the limited variability in these libraries makes it challenging to validate newly discovered or synthesized chemical species [35, 1]. In this context, deep learning-assisted strategies hold the promise to greatly contribute to the prioritization of molecules that should be experimentally evaluated, thus accelerating the rate at which novel drug candidates are found [30].

Using computational methods to estimate and predict properties of molecules, traditionally referred to as quantitative structure–activity relationships (QSAR) modeling, has a long history in material sciences, molecular biology, and biochemistry [28**?**]. The success of these modeling approaches hinges on an appropriate and interpretable representation of the molecules, i.e., chemical descriptors such as the Extended Connectivity Fingerprint (ECFP) [33], and the availability of sufficiently large training data to map structure and "activity". In recent years, the field of computational molecular property prediction has seen major breakthroughs,

largely owing to seminal advances in graph neural networks (GNNs) [10**?** , 16, 45]. These graph-based methods adopt an end-to-end supervised learning framework where predictive models infer a task-specific latent representation from large-scale training data. Important application benchmarks include the Tox21 data [24] where twelve toxic effects of 12,000 environmental chemicals and drugs were measured and provided as prediction challenges, and the MoleculeNet collection [44] which comprises prediction tasks for 16 datasets, divided into four categories: quantum mechanical, physico-chemical, physiological, and biophysical molecule properties. In the context of antimicrobial discovery, a major focus has been the prediction of antimicrobial peptides (AMPs) using curated databases such as CAMP_{R3} (see [40] and references therein). Despite the success of sophisticated deep learning architectures for AMP prediction [26, 19, 36], their task specificity limits the transferability of the learned representation to novel tasks and molecule types. One seminal step toward a general-purpose approach is the use of Directed Message Passing Neural Networks (D-MPNNs) [45] in general antimicrobial discovery [35, 32, 22, 43]. Nonetheless, this approach involved the creation of a custom training set by in-house screening a large number of compounds (ranging from 2,000 to 39,000 molecules) for growth-inhibitory activity against each microbial species of interest. A publicly available large-scale data resource for antimicrobial discovery is thus far elusive, thus prohibiting the straightforward use of end-to-end learning schemes.

In this contribution, we tackle this challenge by introducing a two-stage deep-learning strategy for general molecule prioritization in antimicrobial discovery. The first stage uses a novel self-supervised pre-training strategy, termed `MolE` (`Mol`ecular representation through redundancy reduced `E`mbeddings) for molecule representation. In the second stage, we introduce a set of Antimicrobial Potential scores (APSs) for molecule prioritization that leverage the `MolE` representation and publicly available measurements of growth-inhibiting effects of human-targeted drugs against a diverse set of bacterial species [23]. Inspired by recent molecular self-supervised learning schemes [7, 34, 48], `MolE` uses the large collections of available *unlabeled* chemical structures from PubChem [15] to learn a general-purpose molecular representation that is transferable to arbitrary prediction tasks. Contrary to existing approaches such as MolCLR [48], `MolE` uses Graph Isomorphism Networks (GINs) and introduces the non-contrastive Barlow-Twins pre-training framework [50] to the molecular domain. Furthermore, the `MolE` representation relies purely on unlabeled data and is not informed by any supervised tasks (as, e.g., in [14]). As will be illustrated on a series of curated tasks from MoleculeNet, `MolE`'s pre-trained representation is informative both (i) as a fixed input vector for downstream machine learning and (ii) as a starting configuration for fine-tuned GNN-based prediction. In contrast to huge-scale pre-training schemes such as ChemBERTa [7] or MolFormer [34], the `MolE` framework is computationally light-weight and requires only $\sim$100,000 unlabeled random structures from PubChem to achieve good predictive performance.

Recent evidence has revealed the growth-inhibiting effects of human-targeted drugs against various bacterial species [23, 41, 13]. This behavior can have important consequences for the development of antimicrobial resistance [41, 13], as well as offer interesting candidates for drug re-purposing as novel antimicrobials [35] or adjuvants [18]. Therefore, we combined the `MolE` representation with a recent data resource that measured growth-inhibiting effects of human-targeted drugs against 40 bacterial species [23] to predict the effect that a chemical compound has on the growth of various bacterial strains. Our results showed that the classifier was able to accurately predict the antimicrobial activity of compounds that are structurally diverse from current antibiotics as well as recognize the activity of molecules that were otherwise not recovered by commonly used chemical representations. Furthermore, we used this model to identify non-antibiotic compounds with previously unknown antibacterial activity. Using the output of our model, we prioritized 5 compounds for experimental validation and confirmed the significant effect of 3 drugs on the growth of the human pathogen *Staphylococcus aureus*.

Taken together, our results show that MolE can produce an informative chemical representation. This representation enables widely accessible machine learning algorithms to achieve meaningful performance on property prediction tasks, outperforming baseline and supervised approaches. Furthermore, we have shown that MolE's representation can successfully be applied to the task of antimicrobial discovery. In this way, we envision that methodologies such as ours will allow a wide group of researchers to prioritize molecules for costly experimental evaluation.

Results

MolE pre-training framework

Molecular representation through redundancy reduced Embedding (MolE) is a non-contrastive self-supervised deep learning framework that learns a representation of molecular structures through the application of the Barlow Twin redundancy reduction scheme [50]. The input into our pre-training framework is the text-based representation of chemical structures known as SMILES (simplified molecular-input line-input system) [**?**]. From here, the workflow is detailed in Figure 1a and comprises five main steps: (i) the SMILES are used to construct a molecular graph, where each node represents an individual atom and each edge represents a chemical bond that exists in the molecule, (ii) two augmentations are created for each molecule by masking a subgraph of the original structure, (iii) these augmentations enter a feature extraction step, which is performed by a series of Graph Isomorphism Network (GIN) layers, (iv) the pooled output of each GIN layer is collected to form a final vector representation (r); (v) finally, a non-linear projection head expands each vector representation into an embedding of higher dimensionality and the barlow-twins objective function (L_{BT}) is optimized (see Methods for more details).

By optimizing the L_{BT} objective, the feature extraction step can produce a representation that can aid in the prediction of molecular properties in two ways. First, the static vector representation (r) constructed after the GNN feature extraction step, can serve as input for machine learning algorithms such as XGBoost or Random Forest [4] (Figure 1b). Notably, the dimensionality of MolE's representation ($r \in \mathbb{R}^{1000}$) is comparable to commonly used representations such as ECFP4. Additionally, the parameters of the pre-trained GIN layers can be used to initialize a predictive model based on this GNN architecture, with an MLP placed after the feature extraction step. In this way, the learned GIN weights are updated and fine-tuned to a specific outcome.

MolE enables competitive molecular property prediction

To test MolE's ability to improve the generalization performance of supervised learning schemes, we evaluated the out-of-sample performance of XGBoost and Random Forest when trained with MolE's static features (r), as well as the performance of a GIN-based predictor initialized with MolE's pre-trained weights, on various classification and regression tasks from MoleculeNet [44]. Each dataset was split into training, validation, and testing sets following the scaffold splitting procedure [14], thereby creating a realistic scenario where molecules seen during training are structurally distinct from molecules seen during validation and testing.

We observed that MolE's static representation led to the best performance by XGBoost in the majority of the tasks evaluated (Tables 1 and 2). MolE's XGBoost obtained the best metric in 4 out of 6 classification tasks, with an average increase in performance of 3% with respect to ECFP4. This strategy also achieved competitive, or even greater, performance metrics than the directly supervised D-MPNN algorithm in 4 classification tasks and 2 out of 5 regression tasks. When evaluating fine-tuned models, we observed that MolE achieved the best performance in 3 classification tasks and 4 regression tasks. Notably, the classification tasks where MolE performed best, were among the smallest datasets in our benchmark. Furthermore, MolE's fine-tuned model was able to outperform the D-MPNN in 7 tasks, with an average gain in performance of 3.7% during classification. The added benefit of MolE's static representation is not limited to XGBoost, as we observe similar increases in performance when used in combination with Random Forest predictors (Supplementary Tables 1 and 2). Together, the results obtained in our benchmarking study indicate that MolE produces a molecular representation that is transferable to various property prediction tasks. The static representation enhances the performance of easily accessible and commonly used machine-learning algorithms, while the pre-trained weights improve the out-of-sample performance of GNN-based classifiers. Importantly, both strategies are shown to be able to rival, and even outperform domain-specific supervised learning schemes in D-MPNN, SchNet, and MGCN.

We next performed an ablation study to determine the contribution that different components of our pre-training strategy have on downstream classification performance. We noted that MolE is able to create an informative molecular representation by pre-training on a relatively small dataset of 100,000 unlabeled

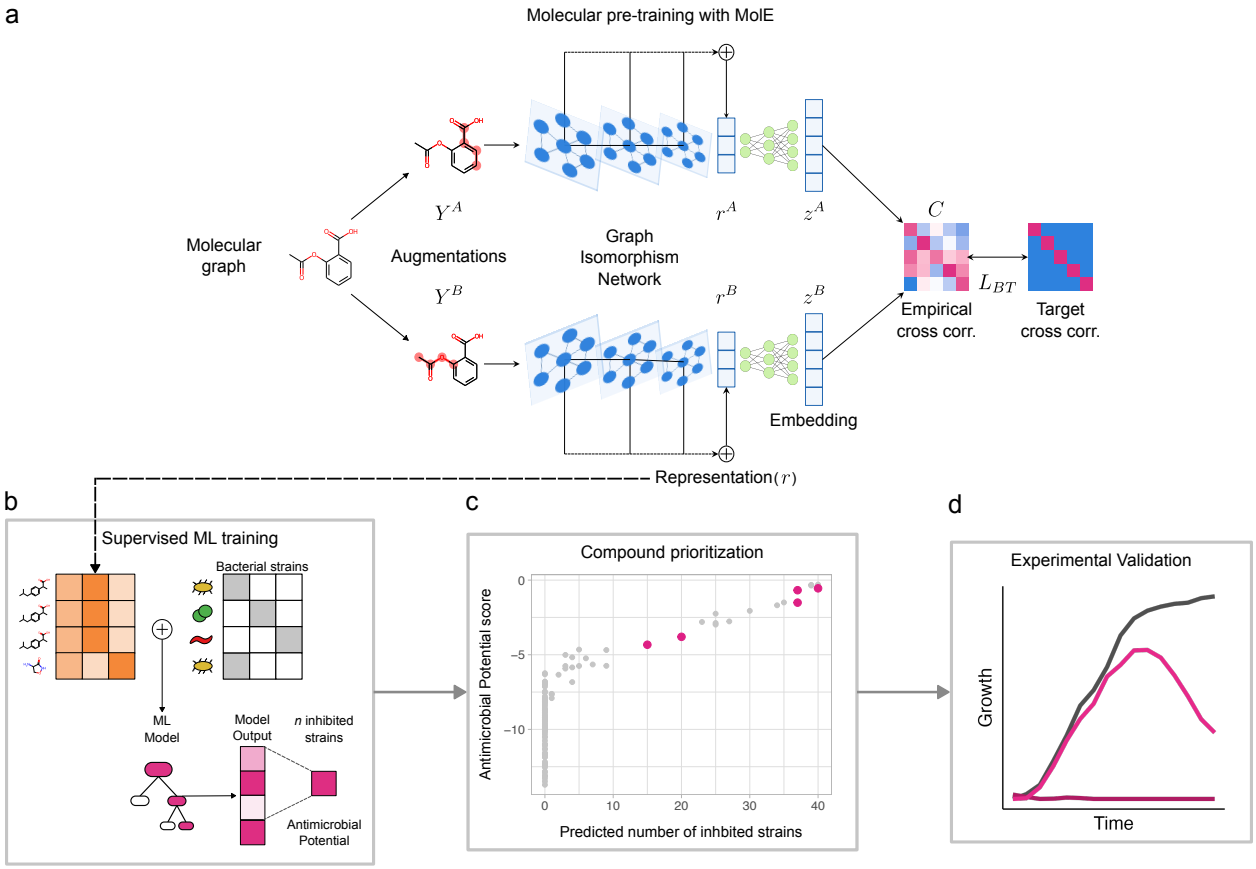

Figure 1: MolE pre-training framework. **a**. For a given molecule, two augmentations (Y^A and Y^B) are created by masking 25% of atoms present and deleting the bonds that exist between them; each augmentation is then encoded by consecutive GIN layers. The update learned at each layer is concatenated to form a final vector representation of the graph (r^A, r^B). An MLP head produces two embedding vectors z^A and z^B which are used to optimize the L_{BT} objective. **b**. After pre-training, the representation (r) is used to train an XGBoost classifier to predict the antimicrobial activity of compounds against various bacterial strains. **c**. The output of this model is used to prioritize compounds for downstream validation. **d**. The selected compounds are screened for antimicrobial activity against human bacterial pathogens.

structures (Supplementary Figure 1d). Further details and results from our ablation study can be found in Supplementary Figure 1 and Methods.

MolE delivers meaningful chemical representations

We further investigated MolE's static features, by examining the similarities between the representation of different molecules. In Figure 2a, we built a UMAP [25] embedding based on MolE's representation of 100,000 molecules that were not used during pre-training. Here, we observe that MolE has learned similar representations for molecules with matching functional groups and/or related topological features. An example of this can be seen on the top, where compounds consisting of a naphthalene group connected to a long carbon chain are placed closely in the embedding space. MolE's representation also recognizes the similarities in the structure of short amino-acid chains, shown in the top left. These small peptides belong to a distinct cluster of molecules in the embedding space.

When compared to ECFP4, we noticed that MolE's representation captures distinct features for the same molecules. As an example, we chose Ractopamine (PubChem ID: 56052, shown in Figure 2c) to serve as our query molecule. We extracted its corresponding MolE and ECFP4 representations and calculated its

Performance on Classification Tasks
(Higher is better)

Dataset	BBBP	Tox21	ClinTox	BACE	SIDER	HIV
# Molecules	2039	7831	1478	1513	1427	41127
# Tasks	1	12	2	1	27	1
GCN [48]	71.8±0.9	70.9±2.6	62.5±2.8	71.6±2.0	53.6±3.2	74.0±3.0
GIN [48]	65.8±4.5	74.0±0.8	58.0±4.4	70.1±5.4	57.3±1.6	**75.3±1.9**
SchNet [48]	84.8±2.2	**77.2±2.3**	71.5±3.7	76.6±1.1	53.9±3.7	70.2±3.4
MGCN [48]	**85.0±6.4**	70.7±1.6	63.4±4.2	73.4±3.0	55.2±1.8	73.8±1.6
D-MPNN [48]	71.2±3.8	68.9±1.3	**90.5±5.3**	**85.3±5.3**	**63.2±2.3**	75.0±2.1
ECFP4	68.51±1.01	70.76±1.48	84.52±0.00	**84.33±1.12**	62.58±1.89	75.77±1.6
N-Gram	74.00±0.01	73.51±0.54	**89.96±1.27**	81.97±0.01	63.68±1.59	**78.89±0.48**
$\text{MolCLR}_{\text{static}}$	65.74±0.68	72.05±1.34	76.04±3.05	68.78±0.00	62.68±2.40	74.29±0.01
$\text{MolE}_{\text{static}}$	**75.98±0.32**	**76.15±0.61**	84.90±0.01	83.84±0.86	**64.48±2.65**	**78.52±0.01**
Hu et.al [48]	70.8±1.5	**78.7±0.4**	78.9±2.4	**85.9±0.7**	62.7±0.8	**80.20±0.9**
HiMol [49]	73.2±0.80	76.2±0.30	80.8±1.4	84.6±0.20	62.5±0.30	74.71±0.23
$\text{MolCLR}_{\text{finetune}}$	73.17±2.13	74.61±1.67	87.79±5.31	81.96±1.11	63.23±3.05	77.61±0.92
$\text{MolE}_{\text{finetune}}$	**75.34±0.93**	75.21±1.77	**92.85±1.33**	83.21±0.93	**64.98±4.42**	78.98±0.57

Table 1: Average ROC-AUC (%) and standard deviation obtained on classification benchmark tasks. The first 5 models are supervised learning methods. The next 4 are the names of molecular features given as input to an XGBoost classifier. The final four methods are fine-tuned models. The best performance metric for each category is marked in **bold**.

Performance on Regression Task
(Lower is better)

Dataset	FreeSolv	ESOL	Lipo	QM7	QM8
# Molecules	642	1128	4200	6830	21786
# Tasks	1	1	1	1	12
GCN [48]	2.87±0.14	1.43±0.05	0.85±0.08	122.7±2.2	0.036±0.001
GIN [48]	2.76±0.18	1.45±0.02	0.85±0.07	124.8±0.7	0.037±0.001
SchNet [48]	3.22±0.76	1.05±0.06	0.91±0.10	**74.2±6.0**	0.020±0.002
MGCN [48]	3.35±0.01	1.27±0.15	1.11±0.04	77.6±4.7	0.022±0.002
D-MPNN [48]	**2.18±0.91**	**0.98±0.26**	**0.65±0.05**	105.8±13.2	**0.014±0.002**
ECFP4	4.08±0.00	1.58±0.02	0.92±0.00	169.85±2.81	0.024±0.000
N-Gram	2.99±0.00	**0.90±0.04**	0.78±0.00	102±2.45	0.027±0.000
$\text{MolCLR}_{\text{static}}$	3.65±0.06	1.43±0.02	0.90±0.02	203.41±0.00	0.030±0.000
$\text{MolE}_{\text{static}}$	**2.78±0.09**	0.95±0.04	**0.73±0.01**	**94.16±6.77**	**0.021±0.000**
Hu et.al [48]	2.83±0.12	1.22±0.02	0.74±0.00	110.2±6.4	0.019±0.000
HiMol [49]	**2.28±0.0**	**0.83±0.0**	**0.70±0.00**	91.5±0.0	**0.018± 0.00**
$\text{MolCLR}_{\text{finetune}}$	3.30±0.12	1.30±0.04	0.74±0.02	69.39±2.00	**0.018±0.004**
$\text{MolE}_{\text{finetune}}$	2.31±0.10	0.90±0.03	**0.70±0.02**	**58.19±1.56**	**0.018±0.000**

Table 2: Average performance and standard deviation obtained on regression benchmark tasks. The first 5 models are supervised learning methods. The next 4 are the names of molecular features given as input to an XGBoost regressor. The final four methods are fine-tuned models. RMSE is shown for FreeSolv, ESOL, and Lipo. MAE is shown for QM7 and QM8. The best performance metric for each category is marked in **bold**.

similarity to all other molecules present in Figure 2a with the cosine and jaccard distances respectively (see Methods). In Figure 2b we observed that while there is general agreement between two representations, discrepancies arise when considering which molecules are most similar to our query (Spearman correlation: 0.42). In Figure 2c it can be observed that different features of our query are highlighted in the molecules determined to be most similar to it by the two representations. While the most highly ranked molecule by MolE shares two phenol groups connected by a carbon chain with one amine functional group with our

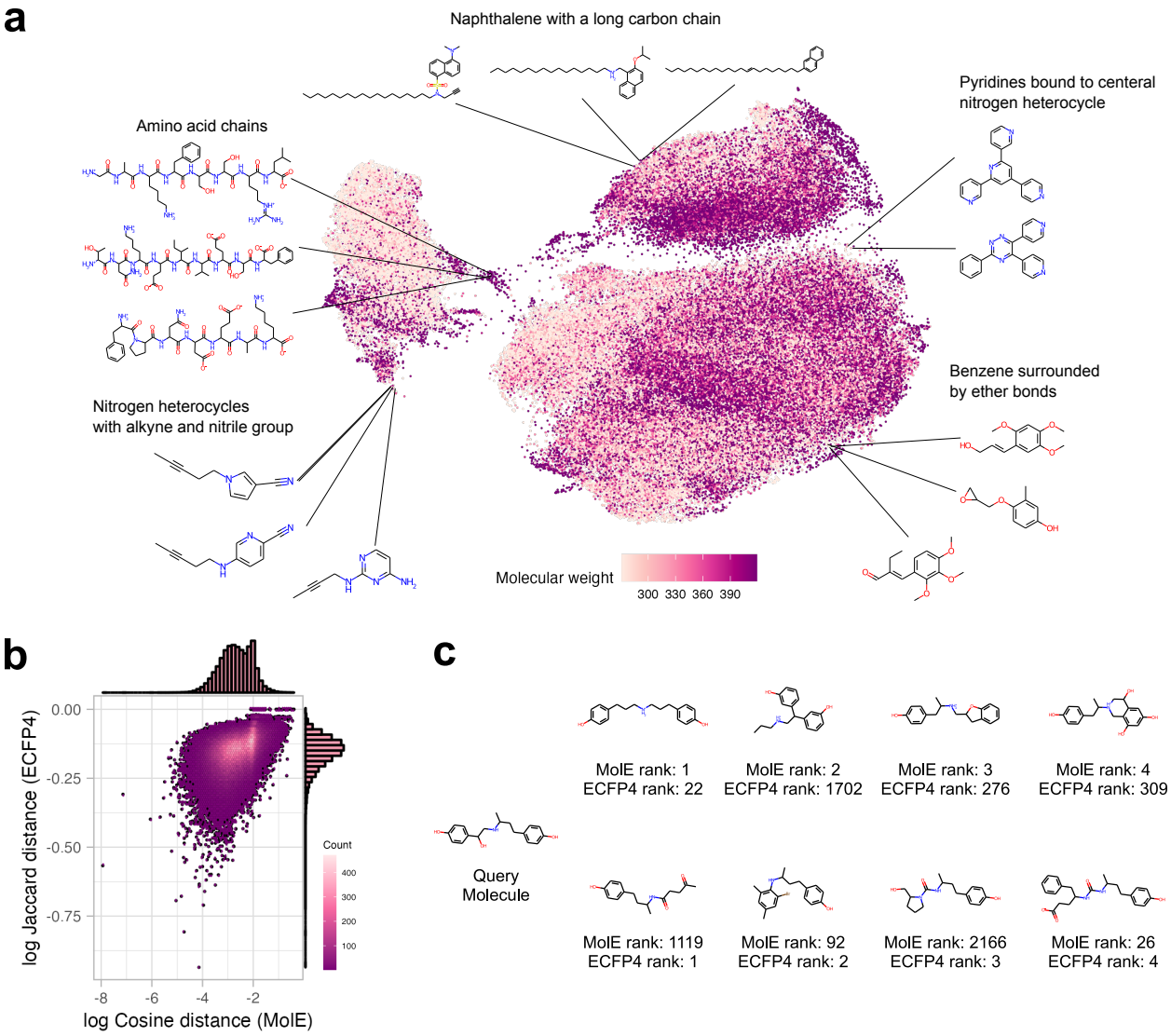

Figure 2: Exploration of MolE's chemical representation. **a**. UMAP embedding based on the cosine distance between MolE's representation of 100,000 chemical structures not used in pre-training. **b**. Comparison of Jaccard distance computed between ECFP4 representations and Cosine distance computed between MolE representations with respect to the query molecule shown in **c**. Comparison of the 4 molecules with the smallest distance to the query according to MolE (top row) and ECFP4 (bottom row)

query, the molecule ranked most highly by ECFP4 only has one phenol ring and a carbon chain with a methyl functional group. Further examples can be seen in Supplementary Figures 2a and b, where MolE was able to highlight global structural features such as a naphthalene group connected to a long carbon chain (Supplementary Figure 2a), as well as the presence of specific functional groups such as sulfonyl chloride (Supplementary Figure 2b).

Together, these observations show that MolE captures chemically relevant information from unlabeled molecular structures. It recognizes the presence of functional groups and learns similar representations for molecules that contain similar structural characteristics. The similarities captured by MolE are related, yet distinct from those uncovered by ECFP4. These differences might explain why our pre-trained representation can improve the performance of machine-learning models for the same given task.

Predicting antimicrobial activity in the microbiome

The discovery of novel antimicrobial compounds is one area in which *in-silico* screening of large chemical libraries has been shown to be useful [35, 20, 32]. Instead of relying on a deep learning-based predictor, we aimed to investigate whether MolE's static representation could enable an easily accessible algorithm, such as XGBoost, to identify novel antimicrobial candidates.

To this end, we made use of the publicly available dataset created by Maier. L., et al [23], which evaluated the effect of 1,197 marketed drugs on the growth of 40 bacterial strains representative of the human gut microbiome (Figure 3a). We trained XGBoost models using one of three possible chemical representations, together with a one-hot encoding of the bacterial strain for which the prediction is made (Figure 3b).In this way, a score is obtained for each compound-microbe combination which indicates the likelihood that the compound will inhibit the growth of the strain in question (Figure 3c). When the score is above a given threshold, then the molecule is predicted to have growth-inhibiting activity. We trained models using MolE, ECFP4, or a collection of chemical descriptive features [1] as input molecular features, and determined the optimal score threshold for each model separately (see Methods, Equation 15. In Figure 3e we observed that XGBoost obtains the best performance on the out-of-sample test set when using MolE's pre-trained representation. The value of this performance metric is also greater than that obtained by Algavi and Borenstein [1] when predicting the effect of new drugs on the same set of microbes (PR-AUC: 73.90).

The added benefit of using MolE's representation can be observed in the concrete example of Diacerein, which was a part of the test set. This drug is used in the treatment of joint diseases and was found to inhibit the growth of 33 microbial strains (Figure 3f bottom row). We found that the model trained with MolE's representation recovers the broad-spectrum activity of this human-targeted medication by correctly predicting 25 of the affected strains (Figure 3f). Notably, the activity of Diacerein is not recovered by the model trained with ECFP4 features. In total, 24 compounds present in the test set were experimentally confirmed to have a broad spectrum of activity, meaning that they inhibit the growth of at least 10 strains [23]. Of these, we observed 5 instances that were not recognized by the model trained with ECFP4, but that were captured by the model trained with MolE's features (Supplementary Figure 4). We observed no instances of compounds with broad-spectrum antimicrobial activity that were recovered using ECFP4 or chemical descriptors and missed by MolE.

Furthermore, the XGBoost model trained with MolE's features was able to re-discover the broad-spectrum activity of Halicin [35] (Figure 3g, bottom row). The experimentally confirmed inhibition of *Escherichia coli* and *Clostridium dificile* was correctly predicted. The predictions made by this model are not trivially positive, as demonstrated by the highly narrow-spectrum activity predicted for Abaucin [20], which was shown to be ineffective against various bacterial strains present in the human microbiome. Our model only predicts growth inhibition against *Eubacterium rectale*, a result which is yet to be experimentally confirmed. These results highlight the fact that MolE's representation enables the widely accessible XGBoost algorithm, to make meaningful predictions of antimicrobial activity. It enables the recognition of the activity of compounds that would otherwise be missed by models using ECFP4 representations and can generalize to molecules that are structurally distinct from known antibiotics, as is the case of Halicin and Abaucin. This is despite the fact that our relatively simple model is trained on a publicly available dataset that contains approximately half of the number of molecules used for their original discovery with D-MPNN.

Discovering novel antimicrobial compounds

With the desire to uncover novel antimicrobial compounds, we applied our XGBoost model trained with MolE's chemical representation, to make predictions on a separate chemical library, consisting of FDA-approved drugs, food homology products, and human endogenous metabolites. We summarised the predicted spectrum of activity of our molecules in two ways: i) as the geometric mean of the scores obtained for all strains, which we call Antimicrobial Potential (see Equation 11) and ii) as the number of strains predicted to be inhibited (Figure 3d, see Equation 14). Predictions were made for a total of 2,327 compounds. Of these, 235 (approximately 10%) were predicted to inhibit 10 strains or more, 77 of which are known antibiotics. The known antibiotics present in our dataset were ranked highly by our model, both in terms of the number of

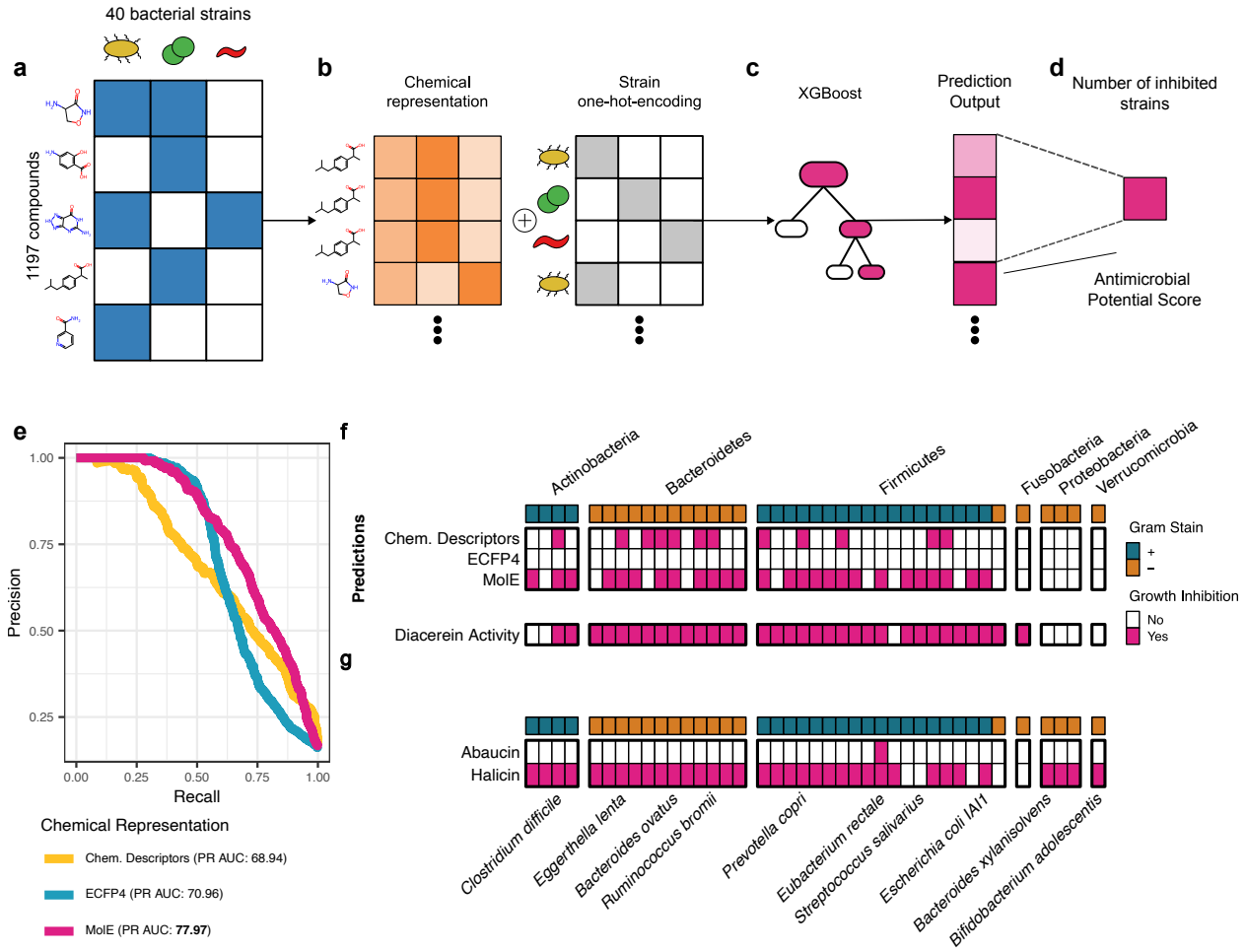

Figure 3: Predicting antimicrobial activity in the human gut microbiome. **a**. We leverage a dataset that determines the antimicrobial activity of 1,197 compounds against 40 bacterial strains [23]. **b**. Compounds are represented as either ECFP4, Chemical Descriptors [1], or MolE's pre-trained representation. This representation is concatenated with a one-hot-encoding vector that indicates the different bacterial strains. **c**. The XGBoost classifier produces a score for each compound-microbe combination, indicating how likely the compound is to inhibit the strain's growth. **d**. The predicted spectrum activity of a compound can be summarised as the total number of strains predicted to be inhibited or as the geometric mean of all scores (Antimicrobial Potential, see Methods). **e** Precision-recall curves were obtained on the test set by models trained with ECFP4 (blue), Chemical Descriptors (yellow), and MolE (pink). PR-AUC is shown in legend. **f**. Each column indicates a strain. The top row indicates gram stain. The following three rows show the predictions of antimicrobial activity for Diacerein made by models trained with Chemical Descriptors, ECFP4, or MolE respectively. The bottom row shows the experimentally validated ground truth [23]. **e** Predictions for the antimicrobial activity of Halicin and Abaucin made by MolE's XGBoost model.

strains they are predicted to inhibit, as well as their Antimicrobial Potential (Figure 4c). Given that we were interested in uncovering molecules with broad-spectrum antimicrobial activity, we considered compounds predicted to inhibit 10 or more strains as our set of positive predictions [23].

As can be seen in Figure 4a, the predicted broad-spectrum antimicrobials include structurally diverse molecules. The model was able to correctly predict the antimicrobial activity of natural products such as Ellagic acid [8] and Shionone [9], as well as the activity of human-targeted drugs such as Doxifluridine [46, 3] and Ospemifene [2].

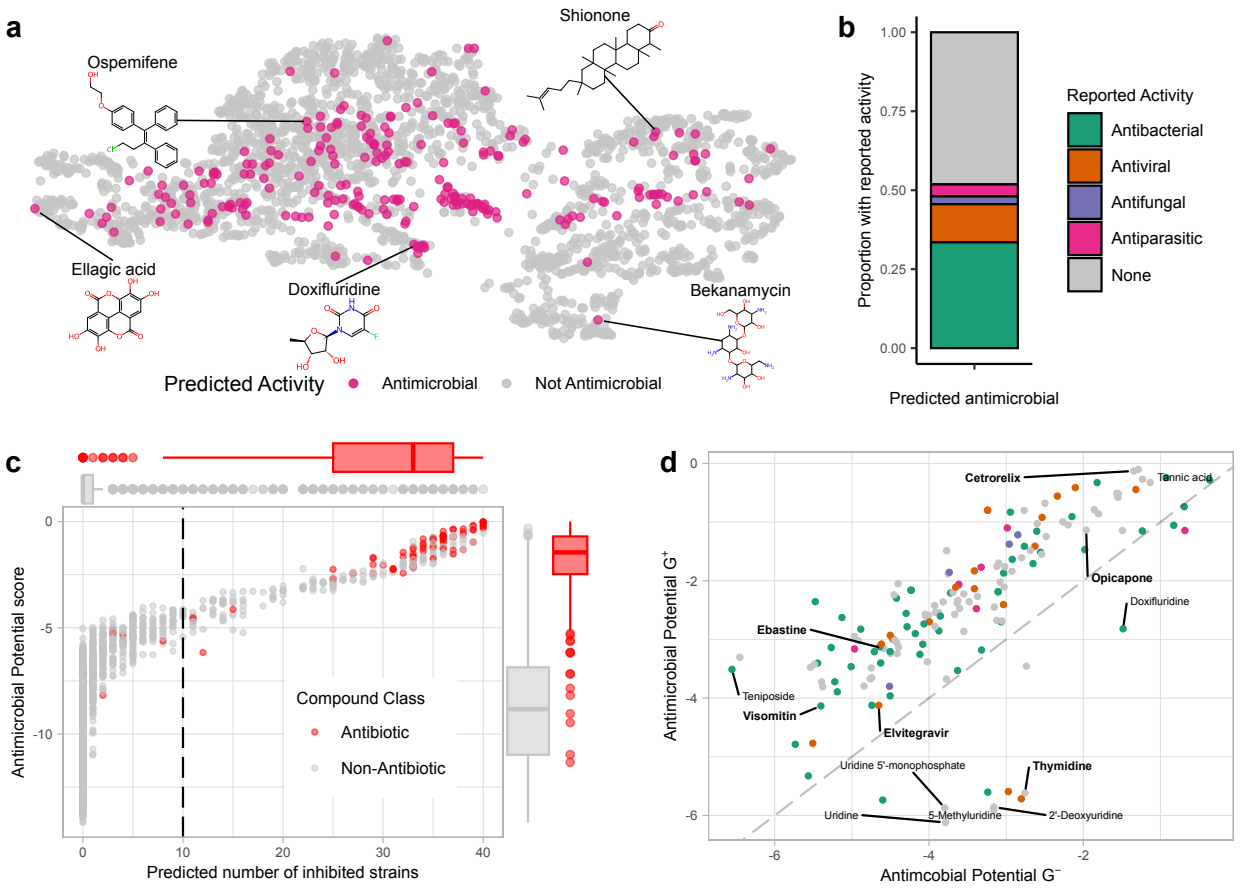

Figure 4: Predicting antimicrobial activity on a new chemical library. **a**. UMAP embedding of MolE's representation of the 2,327 compounds for which predictions are made. Compounds predicted to inhibit 10 or more strains are highlighted in pink. **b**. The proportion of non-antibiotic positive predictions that have been previously shown to have anti-bacterial, viral, fungal, and/or parasitic activity. **c**. Ranking of compounds based on the predicted number of inhibited strains and the Antimicrobial Potential score (See Equation 11). Known antibiotics are shown in red. **d** Comparison of the Antimicrobial Potential scores obtained for gram-positive (G^+) and gram-negative strains (G^-) by non-antibiotic drugs. Names shown in bold were selected for experimental validation.

In total, 158 non-antibiotic compounds were predicted to inhibit at least 10 strains. Of these, 53 (33%) have previously been described by literature to affect the growth of various bacterial species, and a further 29 compounds (18%) are known to have growth-inhibitory activity against viral, fungal, or parasitic species (Figure 4b). The putative 33% hit rate is an improvement over the 1-3% rate commonly cited for large-scale chemical screens [37] and could be further improved by screening the remaining 75 compounds for antimicrobial activity.

We further examined our predictions by comparing the Antimicrobial Potential score obtained for gram-positive strains (A^+) to that obtained for gram-negative strains (A^-), Figure 4d). We note that the majority of the compounds are determined to have a higher Antimicrobial Potential score against gram-positive strains. Notably, among the compounds predicted to be active against gram-negative strains we find uridine and uridine derivatives. Recent evidence has shown that these molecules are powerful adjuvants for the activity of aminoglycosides against *Escherichia coli* [18].

Following our model's predictions, we selected 5 molecules for further experimental validation that had not previously been shown to inhibit the growth of bacterial or fungal strains, that were commercially available,

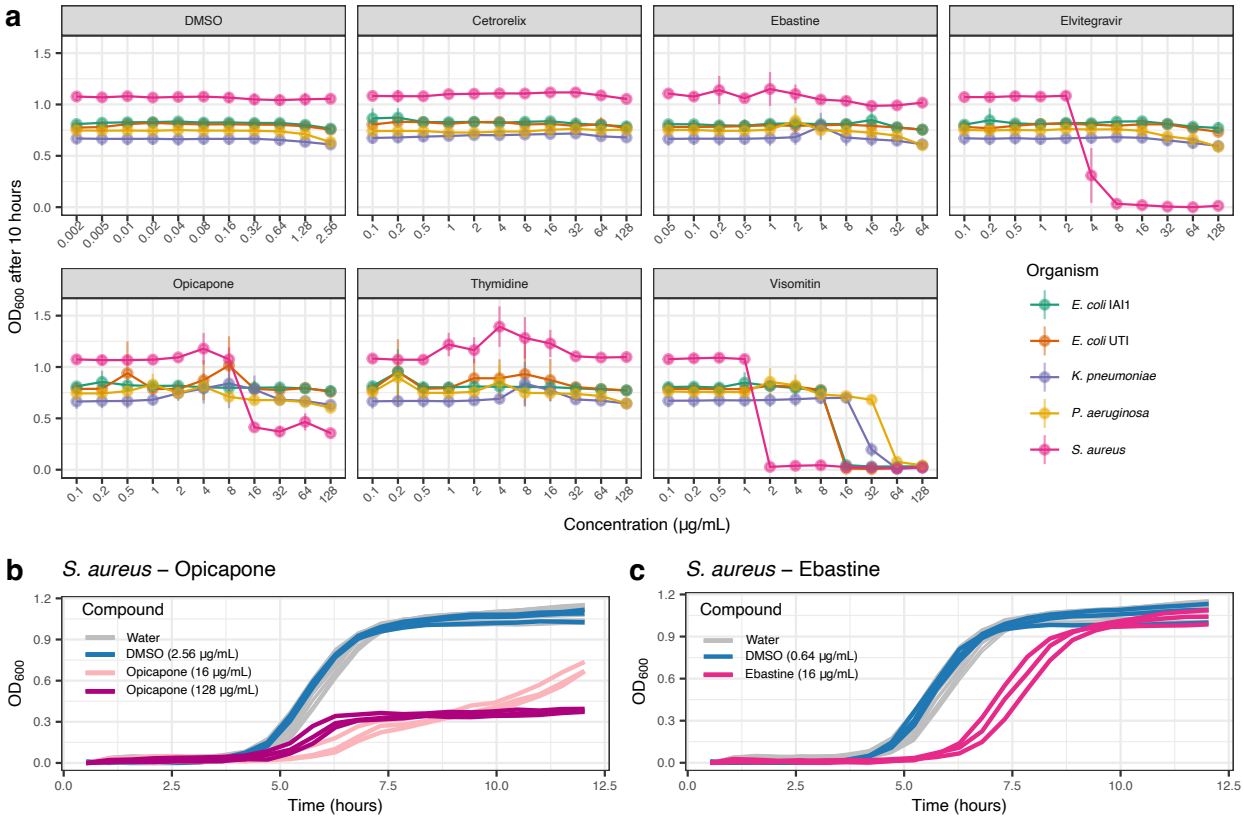

Figure 5: Experimental validation of positive predictions. **a**. Optical density measurements after 10 hours of growth at varying concentrations of each compound. **b**. Growth curve for *S. aureus* when grown in the presence of Water, DMSO, and Opicapone. **c**. Growth curve for *S. aureus* when grown in the presence of Water, DMSO, and Ebastine.

and that covered a variety of functions and chemical structures. The chosen compounds are shown in bold in Figure 4d with more detailed information in Supplementary Table 4. We performed a minimum inhibitory concentration assay against a panel of 5 bacterial strains (Supplementary Table 5).

A summary of the results from our experimental validation can be seen in Figure 5a. We found that Elvitegravir inhibited the growth of *S. aureus* at a concentration of 4 μg/mL. While Opicapone did not completely inhibit the growth of *S. aureus* it did significantly affect its growth dynamics (Figure 5b). At a concentration of 16 μg/mL, this compound extends the duration of the lag phase by approximately 1 hour and increases the population doubling time to 1.43 $\pm$ 0.28 hours from the 0.39 $\pm$ 0.028 hours observed when grown in the presence of DMSO. At the highest concentration tested (128 μg/mL) we observed that the growth of *S. aureus* is limited to a lower carrying capacity, with a maximum optical density measurement of 0.38 $\pm$ 0.01. Additionally, we observed that Ebastine is also able to extend the duration of the lag phase of *S. aureus* by approximately 2 hours (Figure 5c). As a final note, we were able to re-discover the broad-spectrum activity of Visomitin. Our results also showed that this compound can inhibit the growth of *P. aeruginosa* and *K. pneumoniae* at 64 μg/mL, thereby expanding the list of species known to be susceptible to this compound [29]. The complete set of collected growth curves can be found in Supplementary Figures 6 - 11.

The predicted activity of Thymidine and Cetrorelix can be explained by the presence of structurally similar molecules that affect the growth of bacterial strains in the training set. An example of this is Azidothymidine. This medication shares important structural features with Thymidine and was experimentally found to inhibit the growth of 12 strains (9 gram-negative and 3 gram-positive) [23]. This can explain why Thymidine was

predicted to be particularly effective against gram-negative strains (Figure 4d).

We also made predictions on the same set of molecules using the models trained with ECFP4 and Chemical Descriptor features. All models coincided in predicting broad-spectrum antimicrobial activity for Opicapone (Supplementary Figure 5d). The model based on chemical descriptors accurately predicted no antimicrobial activity for Thymidine but also did not recognize the activity of Elvitegravir (Figure **??**c and e). Furthermore, the full set of chemical descriptors was unable to be calculated for some of the compounds present in this chemical library such as Cetrorelix, meaning that this model is limited by the chemical structures for which it can make a prediction (Figure **??**a). The use of ECFP4 features did not recover the activity of our positive control, Visomitin, or the effect that Ebastine has on microbial growth (Figure **??**b and f). Once again, these observations indicate that the use of MolE's features enables predictive models to detect the antimicrobial activity of molecules that would otherwise not be recovered by other commonly used molecular features.

In total, 3 of 5 selected compounds were experimentally confirmed to have varying effects on the growth of *S. aureus*, a noteworthy discovery rate given the small number of molecules screened. These three compounds were predicted to be mainly active against gram-positive strains. While these medications might not become new antibiotic candidates, they are still interesting subjects of study for the stress response mechanisms they might trigger in bacteria. Additionally, the effect that these compounds might have on the human gut microbiome is also an interesting area of study, with potential implications for personalized therapeutic strategies. In this way, MolE enabled the identification of molecules with antimicrobial potential that, at the very least, can enable a better understanding of bacterial mechanisms of survival, and therefore lead to the development of new treatments.

Discussion

In this work, we have presented MolE: a pre-training strategy that learns a novel molecular representation through the redundancy reduction scheme [50] (Figure 1a). Our benchmarking study shows that the representation learned by MolE is transferable to various molecular property prediction tasks. The static representation enhances the out-of-sample performance of widely used algorithms such as Random Forest (Supplementary Tables 1 and 2) and XGBoost (Tables 1 and 2) and consistently outperforms the baseline ECFP4, as well as representations obtained by other self-supervised strategies. It should recognized, however, that ECFP4 is a directly interpretable molecular representation, where each position in the vector indicates the presence or the absence of a specific substructure. Future research can aim to improve the interpretability of MolE's representation thereby enabling researchers to quickly identify interesting structural characteristics. Furthermore, improvements to our current framework could be explored, for example by implementing different types of GNNs, such as graph-attention networks [39] or graph-transformer networks [47].

As a practical application of our framework, we showed that MolE's static features enable XGBoost to make meaningful predictions of the antimicrobial activity of molecules (Figure 3). Indeed, MolE identified the antimicrobial potential of molecules that are missed by ECFP4. MolE also enables the re-discovery of molecules that are structurally distinct from known antibiotics such as Halicin [35], which was originally uncovered using D-MPNN. We highlight the fact that this prediction is made by a relatively accessible algorithm in XGBoost, and relies on a much smaller training dataset. In our current work, we are mainly interested in evaluating the usefulness of our novel chemical representation. Therefore, one limitation of our model for predicting antimicrobial activity is its inability to make predictions for bacterial strains other than the 40 strains studied by Maier et al [23]. Even so, we are able to use these predictions to uncover broad-spectrum antimicrobial activity against bacterial species not present in our training dataset. Future work can seek to incorporate biologically relevant features that describe most bacterial strains, such as those proposed by Algavi and Borenstein [1].

While the experimental screening of chemical libraries is sure to remain a common technique for drug discovery, it is unlikely to keep pace with the continuous introduction of novel chemical structures. Therefore, the use of methodologies like ours is a practical alternative to prioritize molecules for screening, thereby keeping costs and time investment relatively low. As a demonstration of this, we made predictions of antimicrobial

activity on a new chemical library, not used during training or model selection (Figure 4a). We showed that our XGBoost model, trained with MolE's chemical features, is able to recognize the activity of 3 out of 5 compounds selected for experimental validation. These compounds have meaningful and varying effects on the growth of *S. aureus*, including complete inhibition in the case of Elvitegravir, strong restriction on the growth rate and amount of growth in the case of Opicapone, and a lengthening of the lag phase in the case of Ebastine and Opicapone. We are also able to re-discover the activity of Visomitin [29] and further show its activity against other gram-negative pathogens. The screening of additional strains, particularly more gram-positive strains, could further uncover the growth-inhibitory activity of a greater number of our positive predictions, as described by Maier, et al [23]. While these medications might not be re-purposed to be new antibiotic treatments, they are at the very least interesting subjects of study when considering the effect that human-targeted drugs have on microbial life. This is especially relevant, given the growing body of research showing that non-antibiotic drugs contribute to the appearance of bacterial strains resistant to current antibiotics [41, 12, 13, 42]. Future work can dissect the molecular mechanisms that are triggered in bacteria in response to the presence of these compounds, leading to a better understanding of the effect that non-antibiotic drugs have on microbial growth.

Taken together, we believe that MolE addresses a key challenge in the field of drug discovery. By using our pre-trained features, meaningful predictions of various molecular properties can be made by researchers who do not have access to the specialized infrastructure that many deep-learning approaches require. While we highlight the usefulness of MolE for predicting antimicrobial activity, the same representation can be used to prioritize molecules with a different target in mind. In this way, machine-learning methods are poised to complement large screenings and increase the rate at which new treatment candidates are uncovered.

Methods

Pre-training

Dataset. Our pre-training dataset was created by randomly sampling 100,000 unlabeled molecules from a collection of 10,000,000 unique structures originally collected by ChemBERTA [7]. This subset is then randomly split into training (90%) and validation (10%).

Molecular graph construction. Each molecule is represented as a graph G, where all atoms are nodes V and the chemical bonds that exist between them are the edges E. The attributes encoded for each atom are the element it represents and its chirality type (tetrahedral CW, tetrahedral CCW, other). Likewise, each bond is embedded by its type (single, double, triple, aromatic), and its direction (none, end-upright, end-downright).

During pre-training two augmentations (Y^A and Y^B) are created by following the subgraph removal procedure first described in MolCLR [48]. Briefly, a seed atom is selected at random. This seed atom and its neighbors are masked and then neighbors of the neighbors are masked until 25% of the atoms present in the original have been masked. The bonds between these masked atoms are removed, thereby producing a subgraph of the original molecule.

Graph Neural Networks. We explored the ability of GINs and GCNs to extract meaningful features from the molecular graphs constructed in the previous step. Both algorithms make use of an aggregation function to learn an updated representation of each node. In the case of the GIN, aggregation is given as:

$$h_v^l = MLP_{nodes}^l(h_v^{l-1} + \sum_{u \in N(v)} h_u^{l-1} + MLP_{edges}^l(e_{v,u})) \tag{1}$$

The updated representation of node v in the lth GNN layer (h_v^l) is the output of an MLP head (MLP_{nodes}) that receives the representation learned in the previous layer (h_v^{l-1}), the representations of node v's neighbors (h_u^{l-1}) in combination with the output of an MLP (MLP_{edges}) applied to edge attributes between node v and its neighbors u ($e_{v,u}$).

We use summation as our *READOUT* function to obtain a graph-level representation from each layer g^l.

$$g^l = \sum_{v \in V} h_v^l \tag{2}$$

A final graph representation r is obtained by concatenating the result of the summation operation applied to each GNN layer. Here k represents to total number of GNN layers used.

$$r = CONCAT(g^l | l = 1, 2, ..., k) \tag{3}$$

In our setup, the dimensionality of $\mathbf{h_v}$ and $\mathbf{e_{v,u}} \in \mathbb{R}^{200}$, making $g^l \in \mathbb{R}^{200}$ as well. Given that $k = 5$, the final dimensionality of $r \in \mathbb{R}^{1000}$. In experiments where no concatenation is performed, $\mathbf{h_v}$ and $\mathbf{e_{v,u}} \in \mathbb{R}^{1000}$, and r is obtained by applying Equation 2 to g^k. The graph representation learned for augmentation Y^A is correspondingly denoted as r^A, as is the representation learned for augmentation Y^B denoted as r^B.

Non-contrastive learning. Once the final graph representations r^A and r^B are produced, an MLP layer is used to obtain embeddings z^A and $z^B \in \mathbb{R}^d$ respectively. These embedding vectors are used to evaluate the barlow-twins objective function [50]. First, an empirical cross-correlation matrix C is computed between z^A and z^B:

$$C_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b z_{b,i}^A} \sqrt{\sum_b z_{b,j}^B}} \tag{4}$$

Here b indexes the batch samples and i and j index the vector dimensions of the embeddings. In practical terms, $C \in \mathbb{R}^{d \text{ x } d}$, and is the average cross-correlation matrix of a given batch b. Finally, C is used to calculate the barlow-twins loss:

$$L_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{i \neq j} C_{ij}^2 \tag{5}$$

In essence, by minimizing L_{BT} we minimize the difference between C and the identity matrix I. By making $diag(C)$ similar to 1, we ensure that embeddings obtained from augmentations of the same molecule are similar to each other, making them invariant to distortions. At the same time, penalizing off-diagonal values that are different from 0 minimizes the redundancy of the information encoded by the different components of the embedding. The positive constant λ controls the trade-off between these two terms.

Pre-training setup. For all experiments, MolE pre-training is done over 1000 epochs with a batch size of 1000 molecules. An Adam optimizer with weight decay 10^{-5} is used to minimize the L_{BT} loss. The first 10 epochs are performed with a learning rate of 5 x 10^{-4}, after which cosine learning decay is performed. The parameter configuration that minimizes the validation error is kept for downstream tasks.

Benchmarking

Datasets. We collected 11 benchmark datasets originally curated and made available by MoleculeNet [44]. This set included 6 binary classification tasks and 5 regression tasks. In order to make these benchmarking scenarios more challenging and realistic, we separate each dataset into training (80%), validation (10%), and test (10%) sets following the scaffold splitting procedure first described by Hu et.al [14].

Training. For all of our benchmarking experiments, MolE was pre-trained on a set of 100,000 molecules from Pubchem [15]. While different dimensions of the embedding z and values for λ are explored, the dimensionality of the representation is always $r \in \mathbb{R}^{1000}$. Once pre-training is completed, MolE can be used

in two ways: 1) the static vector representation r can be used as input for machine-learning algorithms, or 2) the weights learned by the GNN layers can be used to fine-tune a predictor to a specific task.

In the first case, the pre-trained parameters are not updated any further and are used to encode a given molecule G into a vector representation r. This representation is then used as input molecular features to either a Random Forest or XGBoost. Hyperparameter optimization is performed with a random search over the parameters shown in Tables 3 and 4. Each model configuration was trained three times with different seed values. The model configuration with the largest mean ROC-AUC value on the validation set was then evaluated on the test set.

Random Forest Parameters

Name	Possible Values
criterion	"gini", "entropy", "squared_error"*
max_features	"sqrt", "log2", None
n_estimators	100, 300, 500, 700, 1000
class_weight+	"balanced", None

Table 3: Random Forest parameters considered during a random search. Some parameters are only considered during *regression or +classification tasks

XGBoost Parameters

Name	Possible Values
max_depth	5, 10, 50, 100
eta	0.3, 0.1, 0.05, 1
n_estimators	30, 100, 300, 500, 1000
subsample	0.3, 0.5, 0.8, 1.0

Table 4: XGBoost parameters considered during a random search.

In the case of fine-tuning, an untrained MLP head is placed after the GNN layers. For all classification tasks, the Softmax cross-entropy loss is optimized, while in the case of regression, the L_1 loss is optimized for the QM7 and QM8 tasks and the mean squared error is optimized for all other tasks. A random search is performed over the parameters shown in Table 5. For all tasks, the selected architecture was trained for 100 epochs, using an Adam optimizer with a weight decay of 10^{-6}. While the parameters of both the GNN and the MLP are updated during training, the learning rate chosen for both parts differed. Each model configuration was trained three times.

Finetuning parameters

Name	Possible Values
Batch size	32, 100, 512, 800
Learning rate for MLP head	5 x 10^{-4}, 10^{-3}
N layers in MLP head	1, 2
Activation function MLP	"softplus", "relu"
Learning rate for GNN	5 x 10^{-5}, 10^{-4}, 2 x 10^{-4}, 5 x 10^{-4}
Dropout rate	0, 0.1, 0.3, 0.5

Table 5: Finetuning parameters considered during a random search.

In Table 1, ROC-AUC values for GCN, GIN, SchNet, MGCN, D-MPNN, and Hu et al. are taken from [48]. The ROC-AUC values for HiMol were taken from the original publication [49], except in the case of the HIV task.

Extended Connectivity Fingerprints (ECFP) were calculated using the `GetMorganFingerprintsAsBitVect` function available in RDKit 2020.09.1.0 [17]. In order to get ECFP4 fingerprints we set `fp_radius`=2, and

`fp_bits`=1024.

The $\text{MolCLR}_{\text{GIN}}$ [48] model made available on their GitHub page was used for all benchmarks, following the instructions available in the same repository. The molecular representation obtained after GNN feature extraction was used as the input for either XGBoost or Random Forest. The N-Gram [21] and HiMol [49] models were pre-trained and molecular features were extracted following the default instructions and parameters made available in their respective GitHub repositories.

Ablation study on the MolE framework

In Supplementary Figure 1a, we compared the use of GINs and GCNs in our graph feature extraction step in combination with two alternatives for the construction of the molecular representation r. Concatenated (**C**) representations are obtained as described in Equation 3. Non-concatenated (**NC**) representations consisted of the pooled output of the last GNN layer. In both cases, $r \in \mathbb{R}^{1000}$. We observed that both XGBoost and Random Forest perform best when trained on representations built with the concatenated strategy, independent of the GNN backbone used during pre-training, indicating that each GNN layer captures important information about the molecular structure. Feature extraction performed by GINs and GCNs resulted in comparable performance across most tasks, with the notable exception of ClinTox. In this case, both GIN-derived representations outperformed their GCN counterparts.

We observed increased performance on the ClinTox task when $z \in \mathbb{R}^{8000}$. On all other tasks, the generalizability of the representation produced by MolE is robust to changes in the dimensionality of z (Supplementary Figure 1b).

In Supplementary Figure 1c, we found that the lowest ROC-AUC values are obtained when $\lambda = 10^{-1}$ and increase as this value approaches 10^{-4}. The best XGBoost performance is observed when $\lambda = 10^{-4}$, except in ClinTox and BACE where the highest metric is observed when $\lambda = 5 \text{ x } 10^{-3}$. Finally, in Supplementary Figure 1c, we noted that an increase in the size of the unlabeled dataset used during pre-training does not necessarily lead to improved performance on most classification tasks. A higher ROC-AUC value can be observed in the ClinTox task when MolE's representation is pre-trained on 200,000 molecules.

Exploring MolE representation

The MolE representation (r) obtained after pre-training on 100,000 unlabeled molecules, with $\lambda = 10^{-4}$ and $z \in \mathbb{R}^{8000}$ was used for further investigation. The distance between two MolE representations r^v and r^u was determined using the cosine distance:

$$1 - \frac{r^v \cdot r^u}{||r^v||_2 ||r^u||_2} \tag{6}$$

A UMAP embedding based on the cosine distance between MolE representations was built using the `umap-learn` 0.5.3 Python module.

The distance between two ECFP4 vectors was determined with the Jaccard distance. For two ECFP4 representations u and v the Jaccard distance is given as:

$$1 - \frac{c_{11}}{c_{01} + c_{10} + c_{11}} \tag{7}$$

Where c_{ij} denotes the total number of components where $u = i$ and $v = j$ at the same position.

Compound activity in the microbiome

Dataset. The adjusted p-value table from Maier et al [23] was used to determine labels for the growth-inhibitory effects of the screened compounds. Compound-microbe combinations with an adjusted p-value $<$ 0.05 were considered to be examples of growth inhibition. The 1,197 molecules were divided into training (80%), validation (10%), and test (10%) sets following the scaffold splitting procedure.

Molecular and bacterial strain representation. For a given molecule i we used its corresponding SMILES s_i as input to an encoder function $f(\cdot)$ to obtain the representation m_i:

$$m_i = f(s_i),\ m_i \in \mathbb{R}^d \tag{8}$$

Here, m can be one of three possible representations: 1) ECFP4, in which case $d = 1024$, 2) MolE's representation ($d = 1000$) that was pre-trained on 100,000 unlabeled structures with $\lambda = 10^{-4}$ and $z \in \mathbb{R}^{8000}$, and 3) the chemical descriptors used by Algavi and Borenstein [1] ($d = 92$) which were calculated using RDKit [17].

Let B be the complete set of microbial strains present in the dataset, meaning that $|B| = 40$. We represented each microbial strain $j \in B$ as a one-hot-encoded vector $b_j \in \mathbb{R}^{40}$.

The vectors m_i and b_j are concatenated to form $x_{ij} \in \mathbb{R}^{d+40}$, which is the input to our classifier function $g(\cdot)$. In our work, $g(\cdot)$ is an XGBoost model that, for each x_{ij}, outputs a score y_{ij}.

$$x_{ij} = (m_i, b_j) \tag{9}$$

$$y_{ij} = g(x_{ij}),\ y_{ij} \in \{0, 1\} \tag{10}$$

We summarise the predicted spectrum of activity of compound i two ways: 1) As the Antimicrobial Potential score A_i, which is the geometric mean of the scores obtained for all microbes. We also consider the A_i values obtained on the subsets of gram-positive (A_i^+) or gram-negative (A_i^-) strains:

$$A_i = log((\prod_{j=1}^{40} y_{ij})^{\frac{1}{40}}),\ j \in B \tag{11}$$

$$A_i^+ = log((\prod_{j} y_{ij})^{\frac{1}{|G^+|}}),\ j \in G^+ \tag{12}$$

$$A_i^- = log((\prod_{j} y_{ij})^{\frac{1}{|G^-|}}),\ j \in G^- \tag{13}$$

Where G^+ and G^- are disjoint subsets of B consisting of all gram-positive and gram-negative strains respectively.

2) As the total number of strains predicted to be inhibited K_i:

$$K_i = \sum_{j=1}^{40} h(y_{ij}),\ K_i \in [0, 1, 2, ..., 40] \tag{14}$$

$$h(y_{ij}) = \begin{cases} 1, & y_{ij} \geq p \\ 0, & y_{ij} < p \end{cases} \tag{15}$$

The function $h(\cdot)$ binarizes the output by determining whether the score is greater than a threshold p. We selected p to optimize the F1-Score metric obtained on the validation set for each model (Supplementary Figure 3a). This optimized cutoff was 0.044, 0.068, and 0.209 for the model trained with MolE, Chemical Descriptors, and ECFP4 features respectively.

Model selection and evaluation. A random search over XGBoost parameters mentioned in Table 4 was performed for each chemical representation. The model configuration with the highest PR-AUC on the validation set was then evaluated on the test set.

Prediction of antimicrobial activity

New chemical library. A separate chemical library was constructed based on the FDA-approved Drugs, Human Endogenous Metabolite, and Food-Homology Compound libraries made available by MedChemExpress `https://www.medchemexpress.com/`. The chemical structures for these compounds were gathered from PubChem [15] using the `pubchempy` 1.0.4 module. SMILES were canonicalized and salts were removed using the `SaltRemover` function from RDKit [17].

Compound annotation. Information provided by MedChemExpress included descriptions of the chemicals in the library. The corresponding Anatomical-Therapeutic-Chemical (ATC) code was assigned to each compound by matching compound name strings. A complete collection of ATC codes was taken from `https://github.com/fabkury/atcd`. Overlap with chemicals present in the library used by Maier et al.[23] was determined by matching chemical names and ATC codes. Chemicals present in both libraries were not considered for downstream prediction.

Prediction and evaluation. Molecules with $K_i \geq 10$ were prioritized for further evaluation. We performed a literature search for articles available on PubMed [5] that described the *in-vitro* and/or *in-vivo* antimicrobial activity of our prioritized compounds against any bacterial species. A total of 6 compounds were selected for experimental validation. Criteria considered for compound selection were the following: 1) The compound was predicted to inhibit 10 strains or more, 2) the compound was not an antibiotic, 3) the compound did not have antifungal activity, 4) no previous literature describing the antimicrobial activity of the compound was found, and 5) the compound could be purchased through an independent provider. Furthermore, we attempted to choose compounds with different biological functions that were structurally diverse from each other.

Experimental validation

Bacterial strains and growth conditions. Bacterial strains used in this study are listed in Supplementary Table 5. Prior to the experiments, all strains were cultured overnight in Lysogeny Broth (LB Lennox) adjusted to pH 7.5 at 37°C.

Measurement of Minimum Inhibitory Concentrations (MIC). All compounds were purchased from Biomol GmbH (Germany). Stock solutions were prepared in DMSO and stored at -20^oC till further use. MIC measurements were performed in 96-well plates with 100 μL bacterial cultures in Mueller Hinton (MH) broth using 1:2 serial dilutions of the tested compounds. Starting concentrations of 128 μg/mL were used for Cetrorelix, Opicapone, Thymidine, Visomitin, and Elvitegravir, and 64 µg/mL for Ebastine due to low solubility. No-compound controls contained DMSO. Overnight cultures of three biological replicates of each bacterial strain were adjusted to $OD_{600} = 0.1$ and inoculated into the plates by pinning using a Singer Rotor (Singer Instruments, UK), achieving a 1:200 dilution. Plates were sealed with transparent breathable membranes (Breathe-Easy®, Sigma-Aldrich-Merck) and incubated at 37 °C in a Cytomat 2 incubator (Thermo Scientific) with continuous shaking at 800 rpm. OD_{600} was measured at regular 30-minute intervals for up to 12 h in a Synergy H1 plate reader (Agilent, USA).

Growth curve modelling. All growth curves were normalized by subtracting the minimum of the second, third, and fourth measurements taken. Afterward, each individual curve was modeled as a logistic curve with `Sicegar` 0.2.4 R package. From these curves, parameters such as the maximum growth rate $\mu_{\max}$, end of lag-phase, start of stationary phase, and the carrying capacity were extracted. The maximum doubling time t_{d} was estimated as:

$$t_{\mathrm{d}} = \frac{ln2}{\mu_{\max}} \tag{16}$$

Software and reproducibility

The MolE pre-training framework was implemented using the pytorch-geometric 1.6.3 framework [11] with Python 3.7 [38]. We use the `RandomForestClassifier` and `RandomForestRegressor` implementation available in scikit-learn 1.0.2 [31] and the `XGBClassifier` and `XGBRegressor` objects from xgboost 1.6.2 [6]. The scikit-learn 1.0.2 [31] module is also used when computing ROC-AUC, PR-AUC, and F1 score metrics. The R 4.3.1 language was used for its `ggplot2` 3.4.2 for plotting, `Sicegar` 0.2.4 packages. ECFP4, chemical descriptors, and general SMILES processing were done with the rdkit 2020.09.1.0 package.

Code Availability

The code for the MolE pre-training framework can be found at `https://github.com/rolayoalarcon/MolE`. The steps and models for the prediction of antimicrobial activity are available at `https://github.com/rolayoalarcon/mole_antimicrobial_potential`.

Data Availability

The unlabeled chemical structures used for pre-training were gathered from the MolCLR GitHub repository `https://github.com/yuyangw/MolCLR`. The adjusted p-value table from Maier et. al. [23] was used to train models to predict antimicrobial activity. The chemical library from MedChemExpress can be found in our GitHub repository `https://github.com/rolayoalarcon/mole_antimicrobial_potential`.

References

[1] Yadid M Algavi and Elhanan Borenstein. A data-driven approach for predicting the impact of drugs on the human microbiome. *Nature Communications*, 14(1):3614, 2023.

[2] Stefania Alvisi, Camilla Ceccarani, Claudio Foschi, Maurizio Baldassarre, Alessandra Lami, Marco Severgnini, Tania Camboni, Clarissa Consolandi, Renato Seracchioli, and Maria Cristina Meriggiola. Effect of ospemifene on vaginal microbiome in postmenopausal women with vulvovaginal atrophy. *Menopause*, 30(4):361–369, 2023.

[3] Navid J Ayon and William G Gutheil. Dimensionally enhanced antibacterial library screening. *ACS chemical biology*, 14(12):2887–2894, 2019.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Kathi Canese and Sarah Weis. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1), 2013.

[6] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[7] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[8] Ronita De, Avijit Sarkar, Prachetash Ghosh, Mou Ganguly, Bipul Chandra Karmakar, Dhira Rani Saha, Aniket Halder, Abhijit Chowdhury, and Asish K Mukhopadhyay. Antimicrobial activity of ellagic acid against helicobacter pylori isolates from india and during infections in mice. *Journal of Antimicrobial Chemotherapy*, 73(6):1595–1603, 2018.

[9] Runbao Du, Tian Wang, Hongfa Lv, Yinuo Zou, Xiaoning Hou, Nana Hou, Peng Zhang, Hongen Li, and Gefu Chi. Shionone-targeted pneumolysin to ameliorate acute lung injury induced by streptococcus pneumoniae in vivo and in vitro. *Molecules*, 27(19):6258, 2022.

[10] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

[11] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

[12] Jónína S Guðmundsdóttir, Elizabeth GA Fredheim, Catharina IM Koumans, Joachim Hegstad, Po-Cheng Tang, Dan I Andersson, Ørjan Samuelsen, and Pål J Johnsen. The chemotherapeutic drug methotrexate selects for antibiotic resistance. *EBioMedicine*, 74, 2021.

[13] Suna Sibel Gurpinar, Didem Kart, and Mujde Eryilmaz. The effects of antidepressants fluoxetine, sertraline, and amitriptyline on the development of antibiotic resistance in acinetobacter baumannii. *Archives of Microbiology*, 204(4):230, 2022.

[14] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks, 2020.

[15] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 10 2022.

[16] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[17] G Landrum. Rdkit: open-source cheminformatics, 2006. `https://www.rdkit.org/`.

[18] Manon Lang, Stéphane Renard, Imane El-Meouche, Ariane Amoura, Erick Denamur, Tara Brosschot, Molly Ingersoll, Eric Bacqué, Didier Mazel, and Zeynep Baharoglu. Uridine as a potentiator of aminoglycosides through activation of carbohydrate transporters. *bioRxiv*, 2023.

[19] Chenkai Li, Darcy Sutherland, S Austin Hammond, Chen Yang, Figali Taho, Lauren Bergman, Simon Houston, René L Warren, Titus Wong, Linda MN Hoang, et al. Amplify: attentive deep learning model for discovery of novel antimicrobial peptides effective against who priority pathogens. *BMC genomics*, 23(1):77, 2022.

[20] Gary Liu, Denise B Catacutan, Khushi Rathod, Kyle Swanson, Wengong Jin, Jody C Mohammed, Anush Chiappino-Pepe, Saad A Syed, Meghan Fragis, Kenneth Rachwalski, et al. Deep learning-guided discovery of an antibiotic targeting acinetobacter baumannii. *Nature Chemical Biology*, pages 1–9, 2023.

[21] Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32, 2019.

[22] Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and S Yu Philip. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5879–5900, 2022.

[23] Lisa Maier, Mihaela Pruteanu, Michael Kuhn, Georg Zeller, Anja Telzerow, Exene Erin Anderson, Ana Rita Brochado, Keith Conrad Fernandez, Hitomi Dose, Hirotada Mori, and et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*, 555(7698):623–628, 2018.

[24] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. Deeptox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2016.

[25] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[26] Prabina Kumar Meher, Tanmaya Kumar Sahu, Varsha Saini, and Atmakuri Ramakrishna Rao. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into chou's general pseaac. *Scientific reports*, 7(1):42362, 2017.

[27] Marcus Miethke, Marco Pieroni, Tilmann Weber, Mark Brönstrup, Peter Hammann, Ludovic Halby, Paola B Arimondo, Philippe Glaser, Bertrand Aigle, Helge B Bode, et al. Towards the sustainable discovery and development of new antibiotics. *Nature Reviews Chemistry*, 5(10):726–749, 2021.

[28] Eugene N Muratov, Jürgen Bajorath, Robert P Sheridan, Igor V Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I Oprea, Igor I Baskin, Alexandre Varnek, Adrian Roitberg, et al. Qsar without borders. *Chemical Society Reviews*, 49(11):3525–3564, 2020.

[29] Pavel A Nazarov, Ilya A Osterman, Artem V Tokarchuk, Marina V Karakozova, Galina A Korshunova, Konstantin G Lyamzaev, Maxim V Skulachev, Elena A Kotova, Vladimir P Skulachev, and Yuri N Antonenko. Mitochondria-targeted antioxidants as highly effective antibiotics. *Scientific Reports*, 7(1):1394, 2017.

[30] Mohit Pandey, Michael Fernandez, Francesco Gentile, Olexandr Isayev, Alexander Tropsha, Abraham C Stern, and Artem Cherkasov. The transformational role of gpu computing and deep learning in drug discovery. *Nature Machine Intelligence*, 4(3):211–221, 2022.

[31] 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 E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] ASM Zisanur Rahman, Chengyou Liu, Hunter Sturm, Andrew M Hogan, Rebecca Davis, Pingzhao Hu, and Silvia T Cardona. A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery. *PLOS Computational Biology*, 18(10):e1010613, 2022.

[33] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[34] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

[35] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[36] Paulina Szymczak, Marcin Możejko, Tomasz Grzegorzek, Radosław Jurczak, Marta Bauer, Damian Neubauer, Karol Sikora, Michał Michalski, Jacek Sroka, Piotr Setny, et al. Discovering highly potent antimicrobial peptides with deep generative model hydramp. *Nature Communications*, 14(1):1453, 2023.

[37] Ruben Tommasi, Dean G Brown, Grant K Walkup, John I Manchester, and Alita A Miller. Eskapeing the labyrinth of antibacterial discovery. *Nature reviews Drug discovery*, 14(8):529–542, 2015.

[38] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[39] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

[40] Faiza Hanif Waghu, Ram Shankar Barai, Pratima Gurung, and Susan Idicula-Thomas. Campr3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic acids research*, 44(D1):D1094–D1097, 2016.

[41] Veronica J Wallace, Eric G Sakowski, Sarah P Preheim, and Carsten Prasse. Bacteria exposed to antiviral drugs develop antibiotic cross-resistance and unique resistance profiles. *Communications Biology*, 6(1):837, 2023.

[42] Yue Wang, Ji Lu, Likai Mao, Jie Li, Zhiguo Yuan, Philip L Bond, and Jianhua Guo. Antiepileptic drug carbamazepine promotes horizontal transfer of plasmid-borne multi-antibiotic resistance genes within and across bacterial genera. *The ISME journal*, 13(2):509–522, 2019.

[43] Felix Wong, Erica J Zheng, Jacqueline A Valeri, Nina M Donghia, Melis N Anahtar, Satotaka Omori, Alicia Li, Andres Cubillos-Ruiz, Aarti Krishnan, Wengong Jin, et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, pages 1–9, 2023.

[44] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[45] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[46] Waleed Younis, Ahmed AbdelKhalek, Abdelrahman S Mayhoub, and Mohamed N Seleem. In vitro screening of an fda-approved library against eskape pathogens. *Current pharmaceutical design*, 23(14):2147–2157, 2017.

[47] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. Graph transformer networks, 2020.

[48] Wang Yuyang, Jianren Wang, Cao Zhonglin, and Barati Farimani Amir. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

[49] Xuan Zang, Xianbing Zhao, and Buzhou Tang. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6(1):34, 2023.

[50] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv.org*, Jun 2021.

Acknowledgements

This work was funded by a grant awarded to C.L.M., A.R.B. and C.M.S. for the StressRegNet consortium within the Bavarian research network bayresq.net funded through the Bavarian State Ministry of Science and Arts, Germany.

Author contributions

R.O.A. and C.L.M. conceived the overall objectives and design of the project. R.O.A., C.L.M. and M.R. contributed to the development and evaluation of the pre-training framework. R.O.A., C.L.M. A.R.B., C.M.S. A.Z. and M.K.A. conceived the experimental validation of antimicrobial activity for selected compounds. A.Z. and M.K.A. performed experiments. R.O.A. and M.B. analyzed data from experimental validations. R.O.A. implemented all computational methods and drafted the manuscript. All authors revised and approved the final version of the Article.

Competing interests

The authors declare no competing interests.

Additional information

There is file with supplementary information.

Correspondence

All correspondence and requests for material should be addressed to C.L.M.