JAEGER – Hunting for Antimalarials with Generative Chemistry

William J. Godinez[1,*], Eric J. Ma[2], Alexander T. Chao[1,3], Luying Pei[1,3],

Peter Skewes-Cox[1], Stephen M. Canham[2], Jeremy L. Jenkins[2], Joseph M. Young[1,3],

Eric J. Martin[1], W. Armand Guiguemde[1,3,*]

[1]*Novartis Institutes for BioMedical Research, Emeryville, CA 94608, USA.*

[2]*Novartis Institutes for BioMedical Research, Cambridge, MA 02139, USA.*

[3]*Novartis Institute for Tropical Diseases, Emeryville, CA 94608, USA.*

[*]*Corresponding authors.*

Abstract

Recent advances in generative modeling allow designing novel compounds through deep neural networks. One such neural network model, the Junction Tree Variational Auto-Encoder (JT-VAE), excels at proposing chemically valid structures. Based on JT-VAE, we built a generative modeling approach (JAEGER) for finding novel chemical matter with desired bioactivity. Using JAEGER, we designed compounds to inhibit malaria. To prioritize the compounds for synthesis, we used the in-house Profile-QSAR (pQSAR) program, a massively-multitask bioactivity model based on 12,000 Novartis assays. Based on the pQSAR activity predictions, we selected, synthesized, and experimentally profiled two compounds. Both compounds exhibited low nanomolar activity in a malaria proliferation assay as well as a biochemical assay measuring activity against PI(4)K, which is an essential kinase that regulates intracellular development in malaria. The compounds also showed low activity in a cytotoxicity assay. Our findings show that JAEGER is a viable approach for finding novel active compounds for drug discovery.

Main

Machine learning, and specifically deep learning[1], is poised to drive breakthroughs in multiple disease areas including infectious diseases such as malaria, where the need for novel molecules is as urgent as ever. The parasite-induced disease inflicts ca. 400000 deaths every year, mostly in children under the age of five in sub-Saharan Africa. In the absence of a vaccine, treatments and preventive measures of transmission have been the cornerstone of reducing the burden of malaria. However, current treatments are becoming increasingly less effective against the parasites that have evolved to develop chemoresistance. Drug combinations and a consistent pipeline of new antimalarials will be required to curb the impact of this disease. In this context, machine learning could prove pivotal for the accelerated discovery and development of novel antimalarials[2-5].

The use of computational methods to enhance and accelerate the design of novel active compounds has long been a goal in drug discovery[6,7]. Progress in machine learning, specifically with deep neural networks, allows using these computational models to generate compound structures with desired physicochemical and bioactivity properties. Many of these *generative chemistry*[8-12] models represent molecules through SMILES[13] strings that provide a text encoding of molecular graphs. Because molecule generation is thus rendered as a text generation task, syntax errors lead to invalid SMILES strings that cannot be converted to a molecular structure[14]. Other models represent molecules directly through a weighted graph. Molecule generation in this case is cast as a graph generation task. Graph generation schemes where a single-atom is added at a time[15] (atom-by-atom generation) also lead to chemically invalid intermediate graphs requiring further correction. Models such as the junction tree variational autoencoder (JT-VAE)[16,17] obviate this issue by generating graphs in a sub-structure by sub-structure manner thus consistently yielding valid molecules. The ability of generative chemistry models to consistently propose novel valid molecules is key towards gaining the confidence and uptake of medicinal chemistry teams.

In this work, we developed a JT-VAE-based generative chemistry approach (JAEGER) that couples efficient numerical search strategies with assay activity models to generate novel active molecules. We applied JAEGER to propose novel active inhibitors of malaria. Prioritization of the proposals with Profile-QSAR (pQSAR)[18] models was followed by synthesis and profiling of two compounds. Both compounds exhibited potent anti-malarial activity. Our work thus represents the first report to our knowledge where generative chemistry is successfully used to design novel malaria inhibitors with proven antimalarial activity.

Results

The JT-VAE model represents a molecule through both a junction tree, which represents the arrangement of sub-structures within a molecule, as well as a molecular graph. The tree and graph representations are each mapped onto separate 28-dimensional (28-D) vectors. Within JAEGER, we thus train a JT-VAE model to encode a single molecule onto two 28-D vector representations as well as decode those two vectors back onto a molecule (see **Figure 1a**). The collections of tree and graph vectors corresponding to the molecules in the training set span two 28-D continuous *latent spaces* that JAEGER explores to obtain novel active molecules. Exploration of those latent spaces is guided by an activity model predicting the activity of molecules in the assay of interest, given in terms of the negative logarithm of the half maximal inhibitory concentration (pIC_{50}). The JT-VAE model is optimized jointly for the molecule encoding and decoding tasks as well as the activity prediction task to ensure that the latent representations support all of these tasks (see **Methods**).

Once the JT-VAE model is optimized, JAEGER generates novel active molecules by taking a *seed* molecule as a starting point. JAEGER samples deterministically neighbors around the starting point by defining a number of principal axes around the starting point in both the tree and graph latent 28-D sub-spaces (**Figure 1b**). The direction and magnitude of these principal axes are defined proportionately to the eigenvectors and eigenvalues of the covariance matrices of the tree and graph latent vectors of the molecule training collection, respectively. In both tree and graph sub-spaces, JAEGER samples positions along each axis at intervals defined proportionately to the magnitude of the axis (see **Methods**). Samples in the tree and graph sub-spaces are combined resulting in 56-D vectors that are passed onto the activity model to predict the pIC_{50} values of the joint samples. Only joint samples predicted to have a pIC_{50} above a certain threshold are selected for decoding.

To build the JT-VAE model to design Malaria inhibitors, we used molecules that had been tested in a Novartis-internal *Plasmodium falciparum* proliferation assay (see **Methods**). In total, the dataset had 21065 molecules with measured pIC_{50} values. Once trained, we passed each molecule in the training set through the model and recorded their latent vectors in tree and graph space. To ascertain that the model had learned chemically relevant information, we performed PCA on these vector collections and correlated the resulting principal components projections with known chemical properties, such as molecular weight and the calculated logarithm of the octanol / water partition coefficient (cLogP). Significant correlations were observed especially in the tree space (**Supplementary Figures 1 and 2**), indicating linear exploration of that latent space would yield molecules with varying properties. Correlations between assay activity and individual principal components are weaker, as activity is modeled as function of the entire joint latent space through a residual neural network. Accordingly, activity gradients emerge only over multiple principal components (**Supplementary Figure 3**). Through a random split cross-validation regime (see **Methods**), we also ensured that the learned features supported both tree and graph reconstruction as well as activity prediction. The average tree and graph reconstruction errors for the model were 6% and 8%, respectively, while the correlation of prediction with the experimental pIC_{50} for the latent space activity model was $r^2 = 0.46$. We also ascertained that the model was able to generate valid molecules. We sampled 1000 latent random vectors and decoded them. All latent vectors were successfully mapped onto a corresponding molecule, thus yielding a compound validity of 100% (see **Methods**).

Having ascertained the model's validity, we started the sampling procedure with three proprietary malaria inhibitors as seed molecules. We sampled in total 282 new virtual molecules and calculated their molecular properties (e.g., molecular weight, cLogP, as well as synthetic accessibility scores, SAS[19]). The distributions of those properties are very similar to the distributions computed over the training molecules (see **Figure 2**), thus demonstrating that JAEGER can generate realistic molecules with properties comparable to those tested in the assay. Consensus modeling of activity was achieved by predicting the activity of the sampled molecules with a pQSAR model built for the same *P. falciparum* proliferation assay. The correlation of prediction with experimental values of the pQSAR model was $r^2 = 0.63$ on a random split test set. From the original list of 282 virtual molecules, only the compounds with the top four pQSAR pIC_{50} predictions were selected for synthesis. Among these four, only two compounds (compounds 1 and 2) were synthesized. The properties of these two compounds (**Table 2**) conform well with Lipinski's rule of five[20]. For those two compounds, we computed their Tanimoto similarities to the training set (**Figure 3**). We observed that the bulk of compounds in the training set were substantially dissimilar from the synthesized compounds (mean Tanimoto similarities of 0.18 and 0.17 respectively). We also report the Tanimoto similarities to the seed molecule, as well as to the molecules' nearest neighbors in the training set and the entire Novartis compound archive (**Table 3**). The similarity values are all below 0.67, indicating that these two compounds are different from existing chemical matter.

We investigated whether the proposed molecules that were synthesized had antimalarial activity. Compounds 1 and 2 were tested in vitro against the 3D7 strain of *Plasmodium falciparum*. Upon compound addition and incubation over 72 hours, compounds 1 and 2 were highly active with EC_{50} values of 0.023 μM and 0.025 μM, respectively (**Figure 4a**).

In comparison, approved antimalarial mefloquine used as positive control in the same experiment displayed an EC_{50} of 0.048 μM. Secondly, because the seed molecule from which the two molecules were derived is active against *Plasmodium vivax* PI(4)K (PvPI(4)K), we investigated whether the proposed molecules recapitulated this mechanism of action or inherited a divergent mechanism of action from other molecules in the training set. In a biochemical PvPI(4)K assay, compounds 1 and 2 were active with EC_{50} values of 0.0028 μM and 0.0016 μM, respectively (**Figure 4b**). PvPI(4)K positive control KDU731 displayed an EC_{50} of 172 pM within the same experiment. Lastly, we investigated whether there were potential off-targets associated with cytotoxicity. Compounds 1 and 2 were tested in vitro against HepG2, a hepatocellular carcinoma cell line for liver toxicity. Upon compound addition and incubation over 72 hours, compounds 1 and 2 displayed very low levels of activity with EC_{50} values of 55.83 μM and 60.29 μM respectively, whereas pan-kinase inhibitor staurosporine was highly active with an EC_{50} of 0.09 μM (**Figure 4c**).

Discussion

We developed JAEGER, a generative chemistry approach based on the JT-VAE model. JAEGER includes an efficient sampling technique that deterministically explores the model's latent space to generate novel ideas for compounds. Conceptually, our sampling scheme compares favorably with approaches relying on random sampling[21] from a computation time as well as reproducibility standpoint.

We used JAEGER to design novel antimalarials by first building a JT-VAE model with data from a Novartis-internal malaria proliferation assay. JAEGER yielded quickly realistic novel compound ideas that were well aligned, in terms of their molecular properties as well as synthetic accessibility, with the chemistry explored in the assay. The compounds proposed by JAEGER were also dissimilar to those found in the training set as well as those in the Novartis compound archive. In this sense, JAEGER's sampling technique strikes a balance between existing chemical matter and novel compound ideas. The new compounds 1 and 2 were synthetically accessible in 6 synthetic steps from readily available starting material and displayed reasonable physiochemical properties sufficient to warrant further investigation.

Prioritization of compounds through pQSAR modeling led to the selection of two compounds for synthesis and experimental profiling. The two synthesized compounds exhibited potent antimalarial activity that was on par with approved antimalarials. The compounds also showed potent activity against the *P*. *vivax* PI(4)K kinase and exhibited low cytotoxic activity. While JAEGER overestimated the antimalarial activity, pQSAR's predictions were well within less than half a logarithm of the experimental values, thus emphasizing the need for further in-silico validation of ideas coming out from generative chemistry models with established computational chemistry methodologies.

In conclusion, our study shows the potential of generative chemistry towards the development of novel antimalarials. Application of the JAEGER approach to other disease areas is relatively straightforward and is currently being explored. Further work also involves the modeling of multiple assays as well as ADME parameters to ensure off-target activities and pharmacokinetic parameters are explicitly accounted for. Algorithmic developments to improve computation time are also planned.

Acknowledgments

The authors express their gratitude to colleagues at Novartis that collected the data that was used to build the malaria model. The authors thank Christopher Sarko and Wilian Cortopassi for valuable discussions.

Author contributions

W.J.G. and W.A.G. initiated, designed, and led the study. W.J.G. and E.J. Ma developed and implemented JAEGER. W.J.G. built the malaria model and sampling algorithms. W.A.G. sampled the antimalarial molecule ideas. A.T.C. and L.P. conducted the profiling experiments and collected data. P.S.-C., J.L.J., and S.M.C. provided computational and synthesis resources as well as feedback. J.M.Y. designed the seed compound and provided feedback. E.J. Martin performed cheminformatics modeling and provided feedback. W.J.G., E.J. Martin, and W.A.G. analyzed and interpreted the results. W.J.G. and W.A.G. wrote the manuscript. All authors reviewed the manuscript.

Corresponding authors

Correspondence to W.J.G. and W.A.G.

Competing interests

All authors are (or were at the time of their involvement with the studies) employees of Novartis.

Additional information

Supplementary information for this paper is available at XYZ.

Code availability

The code for JAEGER is available in **Supplementary Code**.

References

1 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).

2 Keshavarzi Arshadi, A., Salem, M., Collins, J., Yuan, J. S. & Chakrabarti, D. DeepMalaria: Artificial Intelligence Driven Discovery of Potent Antiplasmodials. *Front Pharmacol* **10**, 1526, doi:10.3389/fphar.2019.01526 (2019).

3 Lima, M. N. N. *et al.* Integrative Multi-Kinase Approach for the Identification of Potent Antiplasmodial Hits. *Front Chem* **7**, 773, doi:10.3389/fchem.2019.00773 (2019).

4 Bharti, D. R. & Lynn, A. M. QSAR based predictive modeling for anti-malarial molecules. *Bioinformation* **13**, 154-159, doi:10.6026/97320630013154 (2017).

5 Winkler, D. A. Use of Artificial Intelligence and Machine Learning for Discovery of Drugs for Neglected Tropical Diseases. *Front Chem* **9**, 614073, doi:10.3389/fchem.2021.614073 (2021).

6 Rotstein, S. H. & Murcko, M. A. GroupBuild: a fragment-based method for de novo drug design. *J Med Chem* **36**, 1700-1710, doi:10.1021/jm00064a003 (1993).

7 Ertl, P. & Lewis, R. IADE: a system for intelligent automatic design of bioisosteric analogs. *J Comput Aided Mol Des* **26**, 1207-1215, doi:10.1007/s10822-012-9609-3 (2012).

8 Vanhaelen, Q., Lin, Y. C. & Zhavoronkov, A. The Advent of Generative Chemistry. *ACS Med Chem Lett* **11**, 1496-1505, doi:10.1021/acsmedchemlett.0c00088 (2020).

9 Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* **37**, 1038-1040, doi:10.1038/s41587-019-0224-x (2019).

10 Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360-365, doi:10.1126/science.aat2663 (2018).

11 Awale, M., Sirockin, F., Stiefl, N. & Reymond, J. L. Drug Analogs from Fragment-Based Long Short-Term Memory Generative Neural Networks. *J Chem Inf Model* **59**, 1347-1356, doi:10.1021/acs.jcim.8b00902 (2019).

12 Elton, D. C., Boukouvalas, Z., Fugea, M. D. & Chunga, P. W. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering* **4**, 828-849 (2019).

13 Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28**, 31-36 (1988).

14 Li, X. & Fourches, D. SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning. *J Chem Inf Model* **61**, 1560-1569, doi:10.1021/acs.jcim.0c01127 (2021).

15 Liu, Q., Allamanis, M., Brockschmidt, M. & Gaunt, A. L. in *Conference on Neural Information Processing Systems (NeurIPS)* (Montreal, Canada, 2018).

16 Jin, W., Barzilay, R. & Jaakkola, T. S. in *International Conference on Machine Learning (ICML)* (Stockholm, Sweden, 2018).

17 Jin, W., Barzilay, R. & Jaakkola, T. in *International Conference on Machine Learning (ICML)* (Virtual, 2020).

18 Martin, E. J. *et al.* All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration IC50s for 8558 Novartis Assays. *J Chem Inf Model* **59**, 4450-4459, doi:10.1021/acs.jcim.9b00375 (2019).

19 Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform* **1**, 8, doi:10.1186/1758-2946-1-8 (2009).

20 Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**, 3-26, doi:10.1016/s0169-409x(00)00129-0 (2001).

21 Winter, R. *et al.* Efficient multi-objective molecular optimization in a continuous latent space. *Chem Sci* **10**, 8016-8024, doi:10.1039/c9sc01928f (2019).

22 Kingma, D. P. & Welling, M. in *International Conference on Learning Representations (ICLR)* (2014).

23 Kingma, D. P. & Ba, J. in *International Conference on Learning Representations (ICLR)* (Banff, Canada, 2014).

24 Chenouard, N. *et al.* Objective comparison of particle tracking methods. *Nat Methods* **11**, 281-289, doi:10.1038/nmeth.2808 (2014).

25 Godinez, W. J. & Rohr, K. Tracking multiple particles in fluorescence time-lapse microscopy images via probabilistic data association. *IEEE Trans Med Imaging* **34**, 415-432, doi:10.1109/TMI.2014.2359541 (2015).

26 Trager, W. & Jensen, J. B. Human malaria parasites in continuous culture. *Science* **193**, 673-675, doi:10.1126/science.781840 (1976).

27 Johnson, J. D. *et al.* Assessment and continued validation of the malaria SYBR green I-based fluorescence assay for use in malaria drug screening. *Antimicrob Agents Chemother* **51**, 1926-1933, doi:10.1128/AAC.01607-06 (2007).

28 McNamara, C. W. *et al.* Targeting Plasmodium PI(4)K to eliminate malaria. *Nature* **504**, 248-253, doi:10.1038/nature12782 (2013).

Figures

Figure 1. JAEGER overview

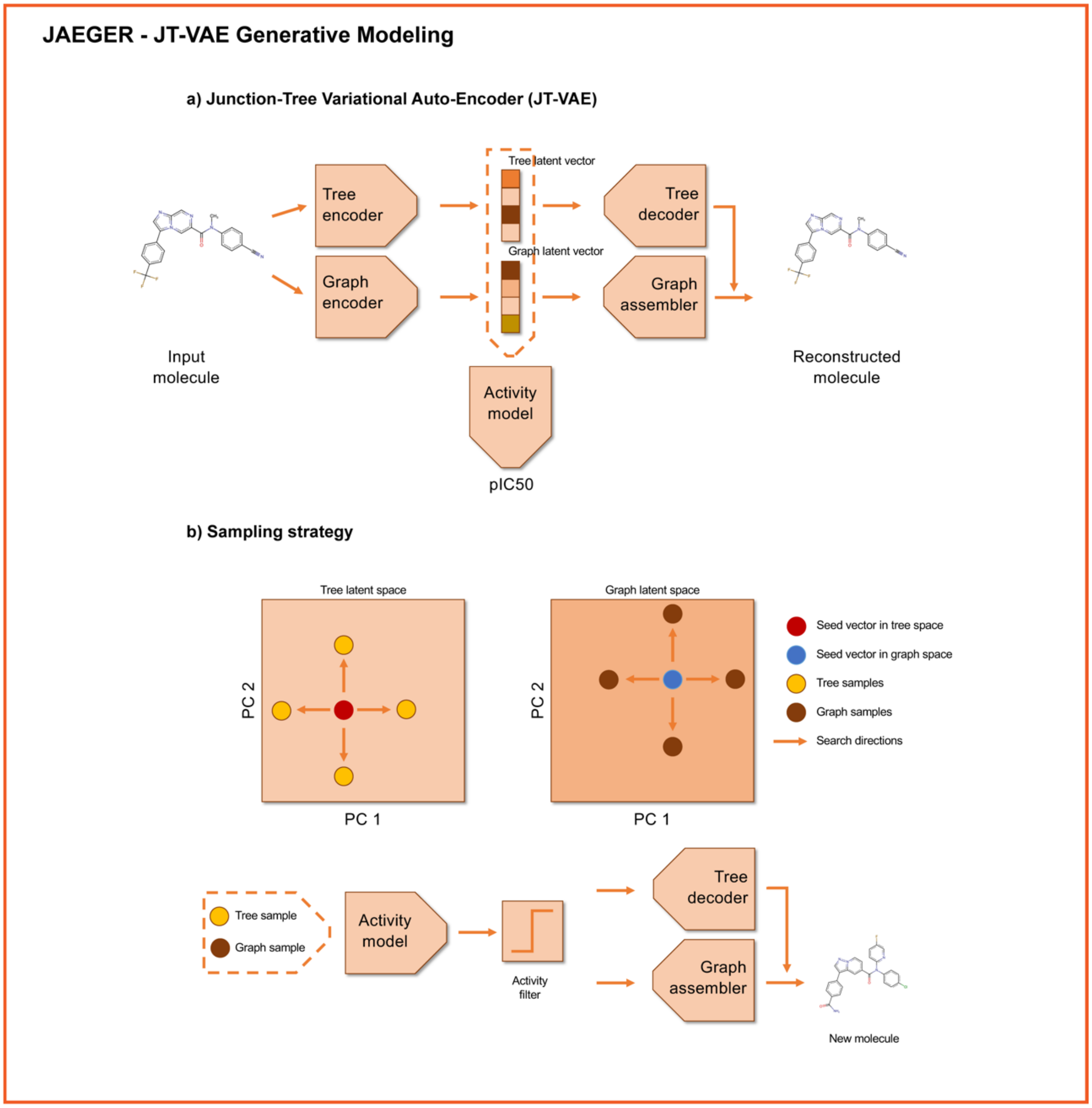


Figure 2. Distribution of calculated properties of training molecules and the 282 molecules generated by the model.

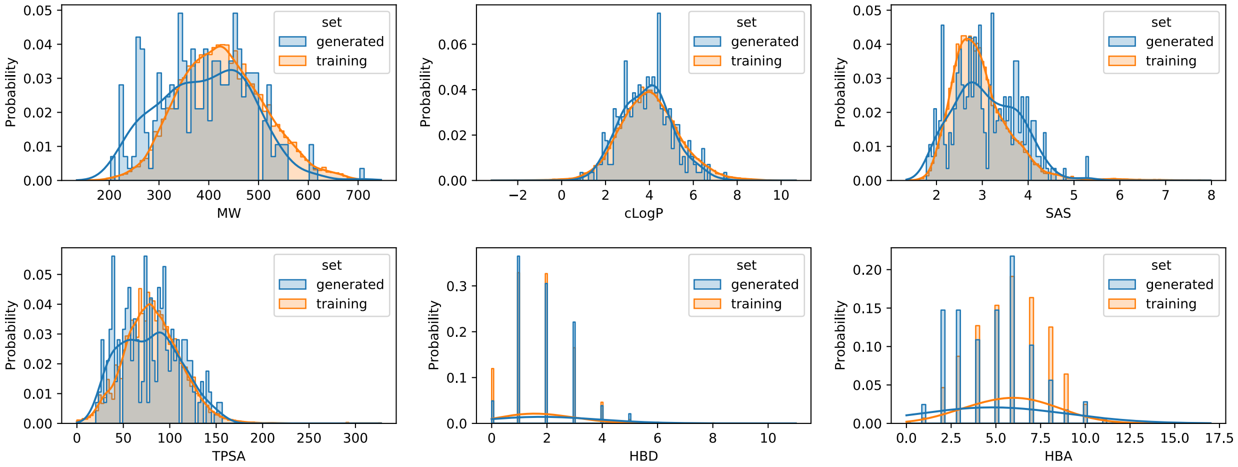
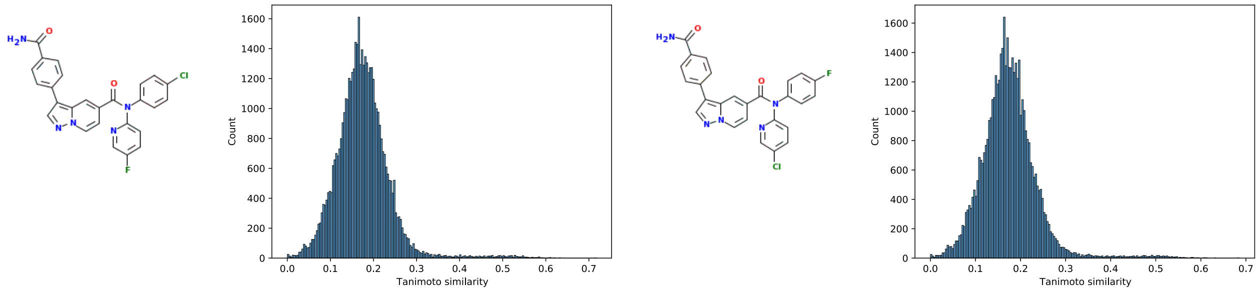


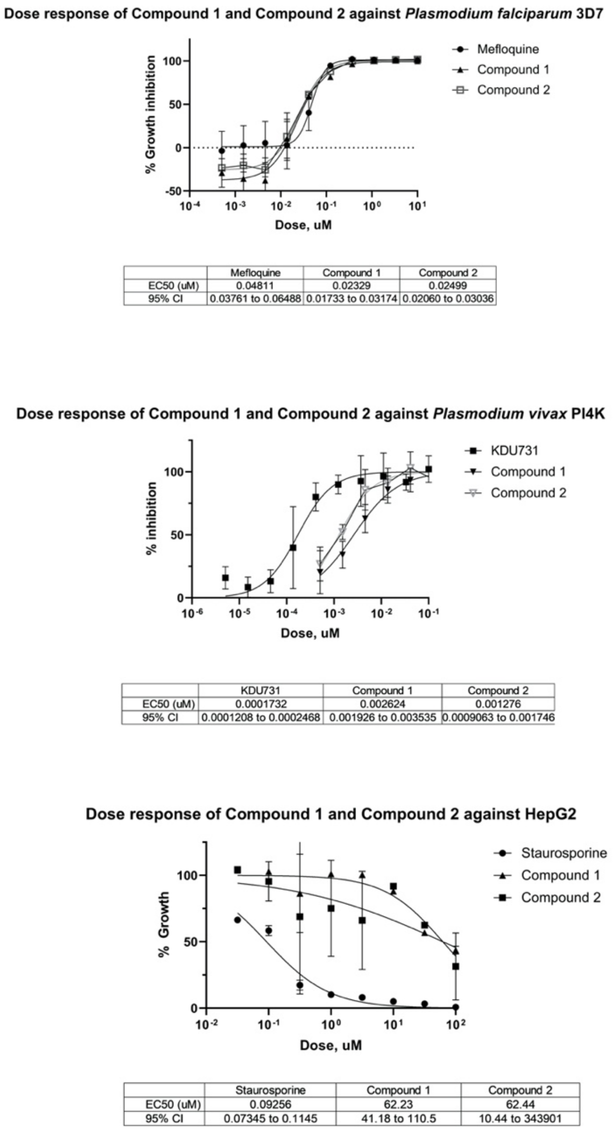
Figure 3. Structures of two synthesized compounds and corresponding distributions of Tanimoto similarities to training molecules.



Compound 1

Compound 2

Figure 4. Profiling results for compounds 1 and 2. 3D7 data were pooled from two independent experiments performed each with two technical replicates. PvPI(4)K data were pooled from two independent experiments performed each with two technical replicates. HepG2 data were pooled from one experiment performed with three technical replicates.



Dose response of Compound 1 and Compound 2 against *Plasmodium falciparum* 3D7

	Mefloquine	Compound 1	Compound 2
EC50 (uM)	0.04811	0.02329	0.02499
95% CI	0.03761 to 0.06488	0.01733 to 0.03174	0.02060 to 0.03036

Dose response of Compound 1 and Compound 2 against *Plasmodium vivax* PI4K

	KDU731	Compound 1	Compound 2
EC50 (uM)	0.0001732	0.002624	0.001276
95% CI	0.0001208 to 0.0002468	0.001926 to 0.003535	0.0009063 to 0.001746

Dose response of Compound 1 and Compound 2 against HepG2

	Staurosporine	Compound 1	Compound 2
EC50 (uM)	0.09256	62.23	62.44
95% CI	0.07345 to 0.1145	41.18 to 110.5	10.44 to 343901

Tables

Table 1. JAEGER model performance metrics

***E*tree**	***E*graph**	**MSE**	**r^2**	**Valid molecules**
6%	8%	0.32	0.46	100%

E_{tree} – Tree reconstruction error

E_{graph} – Graph assembly error

MSE – Mean squared error between experimental and predicted pIC_{50} value

r^2 – squared Pearson correlation coefficient between experimental and predicted pIC_{50} value

Table 2. JAEGER-proposed molecules and calculated properties and activities

Compound	MW	cLogP	SAS	TPSA	HBD	HBA	JAEGER IC_{50} (μM)	pQSAR IC_{50} (μM)
1	485.9	4.14	2.69	93.6	1	5	0.001	0.012
2	485.9	4.10	2.67	93.6	1	5	0.0003	0.017

MW – Molecular weight

TPSA – Topological polar surface area

HBD – Number of hydrogen bond donors

HBA – Number of hydrogen bond acceptors

SAS – Synthetic accessibility score

cLogP – calculated log P

JAEGER IC_{50} – predicted IC_{50} (μM) predicted by JAEGER

pQSAR IC_{50} – predicted IC_{50} (μM) predicted by pQSAR

Table 3. Tanimoto similarities between JAEGER-proposed molecules and other molecules

Compound	Seed molecule	Nearest neighbor in training set	Nearest neighbor in Novartis archive
1	0.4	0.67	0.67
2	0.4	0.65	0.67

Methods

JT-VAE model

JAEGER is based on the Junction Tree Variational Auto-Encoder (JT-VAE) model[16] that represents molecules through both a junction tree subsuming the topology of substructures within a molecule, as well as through a molecular graph capturing the atom and bond structure of a molecule. The JT-VAE model consists of two graph message passing neural networks (MPNNs) that take as input the tree and graph representations, respectively. Each network yields a vectorial representation of the tree and graph $\mathbf{h}_{\text{tree}}$ and $\mathbf{h}_{\text{graph}}$, respectively, that support through a reparameterization trick[22] the calculation of parameters $\boldsymbol{\mu}_{\text{tree}}$, $\boldsymbol{\sigma}_{\text{tree}}$, as well as $\boldsymbol{\mu}_{\text{graph}}$ and $\boldsymbol{\sigma}_{\text{graph}}$ that define the variational posterior approximations for latent variables $\mathbf{z}_{\text{tree}}$ and $\mathbf{z}_{\text{graph}}$. Specifically, the model samples $\mathbf{z}_{\text{tree}}$ and $\mathbf{z}_{\text{graph}}$ from $\mathcal{N}(\boldsymbol{\mu}_{\text{tree}}, \boldsymbol{\sigma}_{\text{tree}})$ and $\mathcal{N}(\boldsymbol{\mu}_{\text{graph}}, \boldsymbol{\sigma}_{\text{graph}})$, respectively. The concatenation of $\mathbf{z}_{\text{tree}}$ and $\mathbf{z}_{\text{graph}}$ serve as input to a residual neural network that predicts the activity, in terms of pIC_{50}, of the input molecule in an assay of interest. To map the latent representations $\mathbf{z}_{\text{tree}}$ and $\mathbf{z}_{\text{graph}}$ back to a molecule, the JT-VAE model adopts a hierarchical approach where first a junction tree is predicted from $\mathbf{z}_{\text{tree}}$, followed by assembly of the molecular graph based on both the predicted junction tree as well as $\mathbf{z}_{\text{graph}}$.

To train the JT-VAE model, we use a set of existing molecules with experimental pIC_{50}s. The goal during training is to improve the performance of the model in terms of its ability to 1) reconstruct accurately both trees and graphs, 2) obtain a good approximation of the variational tree and graph posteriors, and 3) properly predict pIC_{50}s given a latent representation. Performance in those aspects is measured through loss functions 1) L_{tree},

L_{graph} for tree reconstruction and graph assembly cross-entropy errors as defined in[16], 2) L_{KL} for the Kullback-Leibler (KL) divergences between approximated and true posteriors, and 3) L_{mse} for the mean-squared error (MSE) between predicted and experimental pIC_{50} values. An additional loss L_{stereo} tracking the selection of the correct stereoisomer during reconstruction is also included. The loss function L used for parameter optimization is thus given as:

$$L = L_{\mathrm{tree}} + L_{\mathrm{graph}} + \beta L_{\mathrm{KL}} + L_{\mathrm{mse}} + 2 L_{\mathrm{stereo}}$$

where β is a parameter regulating the influence of the L_{KL} term on the optimization process.

For optimization, we use the ADAM algorithm[23] with a base learning rate of 0.003. The learning rate is modulated using an exponential decay function with a multiplicative factor of 0.9. The number of molecules in each training mini-batch is set to 8. Optimization proceeds in two stages: in the first stage, the beta parameter is set to zero. In the second stage, the beta parameter is set to 0.005 and the ADAM optimizer is reset with a base learning rate of 0.0003. Each stage consists of 36 epochs, for a total of 72 optimization epochs. We used the PyTorch implementation provided in the original article[16] as a starting point for our algorithmic developments. We optimized the model parameters using an NVIDIA Tesla K80 GPU with 11.5 GB of memory. For the malaria model, parameter optimization took approximately 135 hours. In this model, we set all hidden layers in the model to include 420 neurons. We set the dimension of both $\mathbf{z}_{\mathrm{tree}}$ and $\mathbf{z}_{\mathrm{graph}}$ to

28. The number of message passing rounds in both tree and graph MPNNs is set to seven. The number of residual blocks in the activity prediction model is set to seven.

Model validation

To validate the model, we evaluate its ability to reconstruct junction trees and molecular graphs as well as to predict the activity of molecules in latent space. We randomly split the molecules tested in the assay into a training set and a hold-out set. The fraction of molecules in the training and hold-out set is 90% and 10%, respectively. We optimize the model's parameters using the molecules in the training set and evaluate the model's performance using the molecules in the hold-out set. For each molecule in the hold-out set, we record whether the model failed to reconstruct correctly the molecules' underlying junction tree as well as molecular graph. We report the fractions E_{tree} and E_{graph} of molecules in the hold out set where failures in junction tree reconstruction and molecular graph assembly occurred, respectively. For each molecule in the hold-out set we also record the predicted activity value. We then compare the predicted and true activity values over all molecules in the hold out set to compute the mean squared error (MSE) as well as the squared Pearson correlation coefficient r^2. **Table 1** show the average values of these errors over the hold-out sets over three random splits.

To check compound validity, we drew 1000 random samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is the mean vector over all 56-D concatenated latent vectors of all molecules in the dataset and $\boldsymbol{\Sigma}$ is the covariance matrix over all 56-D concatenated latent vectors of all molecules in the dataset. The JT-VAE model decoded each sampled vector onto a molecule. The resulting SMILES representation was then converted to a molecule object in the RDKit framework. Compound validity is computed as the fraction of samples for which RDKit successfully converted the decoded SMILES representation onto a molecule object.

Latent space sampling strategy

Once the JT-VAE model is trained, JAEGER generates new active molecules by taking an existing molecule as a starting point. This *seed* molecule is encoded onto its latent representation $\mathbf{z}_{\mathrm{tree}}$ and $\mathbf{z}_{\mathrm{graph}}$. JAEGER samples positions around $\mathbf{z}_{\mathrm{tree}}$ and $\mathbf{z}_{\mathrm{graph}}$ by exploring the neighborhoods around those vectors. Specifically, search axes $\mathbf{e}_{\mathrm{tree},i}$, with $i = 1, 2, \ldots, \mathrm{N}_{\mathrm{axes,tree}}$, are set equal to the eigenvectors of the covariance matrix $\mathbf{S}_{\mathrm{tree}}$ computed over the collection of latent tree representations corresponding to the training molecules. This ensures that JAEGER explores directions providing the largest variations across the training collection. By centering a search axis $\mathbf{e}_{\mathrm{tree},i}$ at $\mathbf{z}_{\mathrm{tree}}$, we can sample $2\mathrm{N}_c + 1$ positions at intervals along the axis:

$$\mathbf{z}_{\mathrm{tree},i,\mathrm{c}} = \mathbf{z}_{\mathrm{tree}} + c\,\sqrt{}\lambda_{\mathrm{tree},i}\,\gamma\,\mathbf{e}_{\mathrm{tree},i}$$

Where $c = -\mathrm{N}_c, -\mathrm{N}_c + 1, \ldots, 0, \ldots, \mathrm{N}_c - 1, \mathrm{N}_c$; where $\lambda_{\mathrm{tree},i}$ is the eigenvalue associated with the eigenvector $\mathbf{e}_{\mathrm{tree},i}$; and where γ is a scaling factor. For a given number of search axes $\mathrm{N}_{\mathrm{axes,tree}}$, the total number of unique neighbors together with the original seed tree latent representation is $\mathrm{N}_{\mathrm{samples,tree}} = \mathrm{N}_{\mathrm{axes,tree}}\,(2\mathrm{N}_c + 1) - \mathrm{N}_{\mathrm{axes,tree}} + 1$. We use the same sampling scheme in the graph latent space to obtain $\mathrm{N}_{\mathrm{samples,graph}}$. By combining all $\mathrm{N}_{\mathrm{samples,tree}}$ tree samples with all $\mathrm{N}_{\mathrm{samples,graph}}$ graph samples, we obtain a total of $\mathrm{N}_{\mathrm{samples}} = \mathrm{N}_{\mathrm{samples,tree}}$

$N_{\text{samples,graph}} - 1$ unique samples in latent space. Each of those 56-D latent samples is passed through the activity model to predict their pIC_{50} values. Only samples with pIC_{50} values above a certain threshold are decoded onto a corresponding molecule. On our hardware, it takes about 2.5 s to decode a latent vector onto a molecule. This deterministic sampling scheme thus ensures that JAEGER is able to generate relatively diverse active molecules while compensating for the relatively slow decoding time of the JT-VAE model. Similar sampling schemes have been successfully used in imaging-related applications[24,25]. For the malaria model, given the explained variances in each sub-space (**Supplementary Figure 4**), we used $N_{\text{axes,tree}} = 28$ tree axes as well as $N_{\text{axes,graph}} = 1$ graph axis. We set N_c to 25 and γ to 0.1 to sample densely around the seed molecule.

Dataset

We used molecules from a Novartis-internal *Plasmodium falciparum* proliferation assay. Compounds with more than 50 atoms were left out from the dataset in order to minimize reconstruction errors with larger molecules. Molecules with activity measurements outside the dose range were left out from the training set as well. After filtering, the dataset had 21065 molecules with measured pIC_{50} values. Properties of the molecules were calculated with RDKit.

Plasmodium falciparum **proliferation assay**

Following an established *P. falciparum* protocol[26], parasite cultures were grown with complete media (RPMI 1640 medium (10.4 g/l) with 0.5% AlbuMAX II, 200 μM hypoxanthine, 50 mg/L gentamicin sulphate, 35 mM HEPES, 2.0 g/L sodium bicarbonate and 11 mM glucose) and human erythrocytes. Cultures were maintained at 37° C in an incubator with 5% CO_2. In vitro antimalarial activity was measured according to a modified SYBR Green cell proliferation assay[27]. Dose response curves data were normalized based on fluorescence signal values from DMSO treated wells (0% inhibition) and mefloquine treated wells (100% inhibition) at a final concentration of 10 µM. The standard logistic regression model was applied for curve fitting in order to determine EC_{50} (GraphPad Prism Software), based on two independent experiments each performed in duplicate.

PI(4)K enzymatic assay

Enzymatic activity was measured according to a previously reported method[28]. Briefly, L-α-phosphatidylinositol (Avanti Polar Lipid), dissolved in 3% *n*-octylglucoside (Roche Diagnostics), was used as the lipid substrate for the PI(4)K activity assay. PvPI(4)K was assayed using Transcreener ADP_2 FP detection kit (BellBrook) in a black, solid 384-well plate (Corning). EC_{50} values were calculated from two independent experiments performed in duplicate using GraphPad Prism software.

HepG2 cytotoxicity assay

HepG2 are adherent cells that were maintained in Dulbecco's Modified Eagle's Medium (DMEM)-F12 supplemented with 1% penicillin/streptomycin and 10% heat-inactivated FBS. Fifty microliters of a $5x10^4$ cells/mL suspension were dispensed in white, solid 384-well plates (Greiner). Cells were exposed to compounds for 72 hours, after which cell viability was quantified by using CellTiter-Glo®. This reagent measures ATP release based on the mono-oxygenation of luciferin catalyzed by Mg^{2+}, ATP and molecular oxygen. EC_{50} values were calculated from two independent experiments performed in triplicate using GraphPad Prism software.