VARIATIONAL AI

Generative AI Technology

Generative Machine Learning Can Construct Smooth Chemical Search Spaces for Efficient Drug Discovery

2022

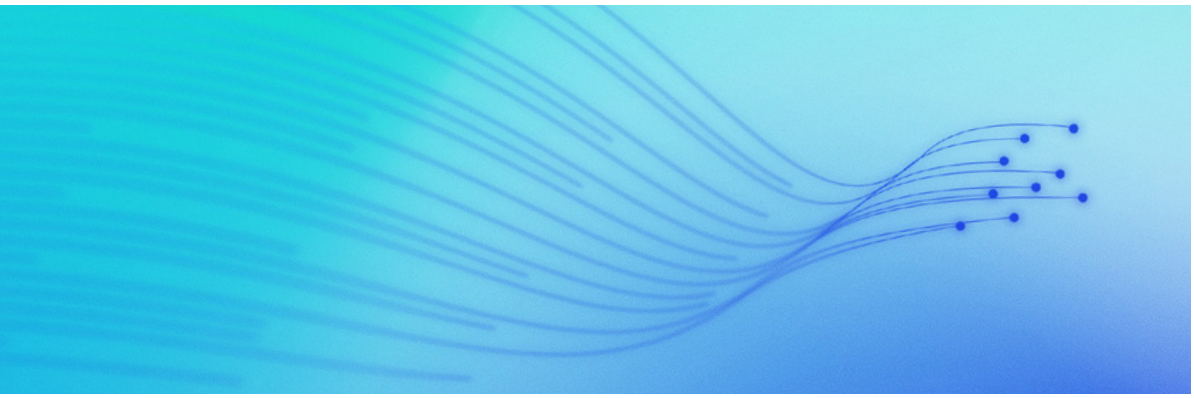
Introduction

The fundamental steps of rational drug design include the identification of a clinically relevant target protein, the discovery of "hit" ligands that weakly modulate the target protein in the desired manner, and the optimization of selected hits for high potency against the target, low potency against all related off-targets, and good absorption, distribution, metabolism, excretion, and toxicity (ADMET). Target identification and the optimization of selected hits are informed by biology and medicinal chemistry, respectively. In contrast, the initial discovery of hits that are active against the target, and can be effectively optimized, is more dependent on exhaustive searching and luck. Conventional approaches for hit discovery are expensive, inaccurate, and can only explore a minuscule fraction of the full space of synthesizable, drug-like molecules.

Generative machine learning (ML) promises to efficiently optimize more accurate estimates of binding affinity and other pharmacological properties over the entirety of drug-like chemical space. Rather than exhaustively testing a screening library or making small changes to the best known compounds, generative ML maps chemical space to a smooth search space in which small moves correspond to small changes in potency, ADMET, etc. Within this search space, a diverse set of potent, selective, lead-like, and novel hits can be found efficiently. These hits can be drawn from the full extent of make-on-demand compound libraries comprising tens of billions of molecules, or even from de novo synthetic space. Hits can be jointly optimized for many properties simultaneously, making subsequent lead optimization easier, faster, and less expensive.

The Limits Of (Virtual) High Throughput Screening

Drug hunters often identify novel initial hits by testing large libraries of compounds, either experimentally via high throughput screening (HTS); or computationally via virtual high throughput screening (VHTS).[1] HTS is expensive, and costs increase in proportion to the size of the screening library. While tens of billions of compounds are available for purchase, HTS is generally limited to between thousands and millions of compounds. VHTS expands the fraction of chemical space searched while minimizing experimental costs by computationally predicting binding affinity on a fixed library of drug-like molecules, and selecting only a small fraction with desirable predicted properties for further wet lab investigation.



VHTS may be performed using a ligand-based or structure-based pharmacophore. Such pharmacophores comprise significant ligand-protein interactions (hydrogen bonds, ionic, hydrophobic, etc) that are predicted to be consistent amongst strong binders (yielding a ligand-based pharmacophore) or are predicted to be consistent with the protein target (producing a structure-based pharmacophore). A chemical library can then be screened for compounds with low-energy conformers that match these pharmacophoric interactions, while avoiding obvious steric clashes with the target protein. Pharmacophores sidestep any mechanistic simulation of the dynamics of ligand-protein binding, maximizing computational efficiency, but their disregard for the physics of binding sharply limits their ability to generalize.

Binding affinity for VHTS may also be predicted using molecular docking, which uses the 3D structure of the target protein and knowledge of ligand-protein interactions to infer favorable binding modes of the ligand via a semi-heuristic minimization of the free energy. The resulting estimate more accurately captures the physics of binding than a pharmacophore model. Nevertheless, the free energy of binding is strongly affected by reorganization of the network of water molecules surrounding the ligand and protein, the deformation of the target protein induced by the binding of the ligand, and the loss of configurational entropy upon binding. These phenomena are difficult to capture accurately without a series of explicit solvent molecular dynamics (MD) simulations of every atom over time, requiring overwhelming computation.

Novel chemistries must be explored to discover new classes of compounds with superior properties, and to avoid existing patents. Fortunately, the chemical space available for drug discovery is astronomically large.

Novel chemistries must be explored to discover new classes of compounds with superior properties, and to avoid existing patents. Fortunately, the chemical space available for drug discovery is astronomically large. Commercial make-on-demand libraries include tens of billions of compounds, and are growing at an exponential rate. De novo synthesis can easily access many orders of magnitude more compounds, and the full space of drug-like molecules is estimated to be between 1020 and 1060.[2,3]

On the other hand, the cost of HTS and VHTS scale linearly with library size. Even molecular docking cannot be run exhaustively on increasingly large make-on-demand libraries, let alone the full space of synthetically accessible, drug-like molecules. Alternative approaches to efficiently optimize over large chemical spaces, based upon more accurate estimates of experimental binding affinity and other molecular properties, are required to leverage progress in synthetic techniques for hit discovery.

Deep Learning Promises to Improve Molecular Property Prediction

Rather than explicitly modeling the ligand-protein binding process from first principles, it is possible to take a data-driven approach. Ligand- and structure-based pharmacophores are a simple example of this in which the activity of a query ligand is predicted based upon overlap with the pharmacophore of known actives. Quantitative structure-activity relationship (QSAR) models utilize data more flexibly, by applying simple machine learning (ML) techniques (typically decision trees, support vector machines, or neural networks) to features such as physicochemical properties or molecular fragments, to predict experimental activity.[4]

In general, these data-driven techniques define a rule that separates actives from inactives based upon a set of trainable parameters. For a ligand-based pharmacophore, the rule may be a threshold for the intersection-over-union of the pharmacophoric features of the query ligand versus the target pharmacophore. QSAR models may divide actives from inactives using a hyperplane in the space defined by the input features (linear models), a hyperplane in a space defined by nonlinear transformations of the input features (support vector machine or neural network), or a succession of thresholds on individual input features (decision trees).

Since around 2009, a deep learning revolution has transformed ML.[5] Whereas conventional machine learning algorithms used networks with one or two sequential layers of processing on carefully engineered features, deep learning algorithms have grown to thousands of sequential layers, hundreds of billions of parameters, and learn their own input features directly from the data. Such powerful deep learning models, with more complicated rules separating actives from inactives, promise to support more accurate activity prediction. However, they must generalize far beyond the experimental data used to train the models.

The cheminformatics community generally believes QSAR models can interpolate between data points, but not extrapolate outside of the available data, and thus cannot be used to perform VHTS over large, diverse libraries. Considerable effort has been devoted to characterizing the applicability domains of QSAR models based upon structural or physicochemical fingerprints, within which models are expected to interpolate accurately.[6]

In contrast, deep learning algorithms extrapolate accurately in domains like images, text, and speech. Small, semantically irrelevant changes in images, including shifts, scales, and rotations, induce enormous changes in the set of pixels that define an image. As shown in Figure 1, the pixel-based representations of images that depict the same class of object (eg, a cat) are almost as far from each other as they are from images that depict different object classes (eg, dogs, cars, houses). As a result, image classification in pixel space requires extrapolation from distant data points of the same class, rather than interpolation from nearby data points from different classes.

Deep learning algorithms nevertheless achieve excellent classification performance when trained on datasets comparable in size to those available for drug discovery. The most popular image dataset for machine learning is probably ImageNet, which has ~1,000 "actives" for each of 1,000 classes. State-of-the-art deep learning algorithms predict the correct class, out of 1,000 possibilities, on over 88% of test images; comparable to the 95% accuracy achieved by humans.[7,8] This exceptional performance is only possible because deep learning algorithms construct a representation that is smooth with respect to image semantics, rather than pixels, as shown in Figure 1.

Figure 1 – 2d embedding (t-SNE) of images (left) and their labels (right) based upon their pixel representation (A) and representations constructed by a deep learning algorithm (B). In the native pixel representation, nearest neighbors are rarely of the same class, and extrapolation is difficult. The deep learning representation naturally organizes images by class, and extrapolation is easy.

The size of ImageNet is analogous to BindingDB, which has over 1M activity data points on over 500k ligands and thousands of protein targets.[9] Carefully designed deep learning algorithms thus promise to significantly improve the accuracy of molecular property prediction and optimization. To realize their full potential, such algorithms must complement the structure of the input just as image recognition

algorithms embody the invariance of image content to small shifts, and address the noise endemic to pharmacological datasets.

Recently, ML has achieved notable successes in domains closely related to drug discovery. AlphaFold uses ML to predict the 3D structure of a protein based upon sequence and structure information from a large database of homologous proteins.[10] Fragments of 3D structure from similar proteins are nonlinearly stitched together, with residue sequences that exhibit complementary mutations across many homologous proteins encouraged to remain bound together. AlphaFold's output corresponds to a holo structure. It does not capture the change in protein conformation induced by a particular small molecule, and cannot be used directly to predict ligand-protein interactions. Nevertheless, complementary applications of ML promise to infer the relationship between structure and properties of small-molecule ligands.

Generative Ml Searches Efficiently Over Chemical Space

Small changes in molecular structure can induce large changes in molecular properties. This phenomenon, particularly evident across activity cliffs, makes optimization difficult and inefficient.[11] The sort of hill climbing performed by medicinal chemists during lead optimization, in which small modifications of the current best molecule are evaluated experimentally and the best modification is used as the basis for the next round of optimization, will generally get stuck in a bad local optimum. These local optima are like foothills surrounding a large mountain. A path that goes directly uphill from a random starting point usually gets trapped in one of these foothills. There is no sequence of small, beneficial modifications from such a bad local optimum to the best possible compound.

This problem cannot be fully solved by using (V)HTS to find hits from which to begin lead optimization. While (V)HTS searches thousands or millions of starting points, chemical space is so vast, with as many as 1060 molecules, that no compound in a fixed screening library is likely to lie near the global optimum. Moreover, it is all but impossible to predict the quality of the final optimized lead from the starting hit considered by (V)HTS, and only a handful of (V)HTS hits can be subject to lead optimization.

Conventional drug discovery approaches are like searching for an address in a phone book, where business names are in alphabetical order but geographical location is unordered.

Rather than searching within the native space of molecular structures, generative ML transforms to a learned search space where nearby molecules have similar properties (binding affinity, ADMET, etc), but are less structurally similar, as shown in Figure 2. Conventional drug discovery approaches are like searching for an address in a phone book, where business names are in alphabetical order but geographical location is unordered. Generative ML is like searching for an address on a map, where such spatially-based exploration is intuitive and efficient.

Generative ML constructs this smooth, natural search space by learning a pair of mappings: an encoder from molecules to the search space, and a decoder from the search space back to the space of molecules. The mappings are trained so that molecular properties, such as binding affinity and ADMET, change smoothly with respect to, and can be predicted easily and accurately from, the position in the search space. When properly formulated, with a term encouraging the encoder from molecules to search space to contain no extraneous information, this algorithm is called a variational autoencoder.[12] Deep learning algorithms, such as deep neural networks, are used for the encoder and decoder of such variational autoencoders, to maximize performance.

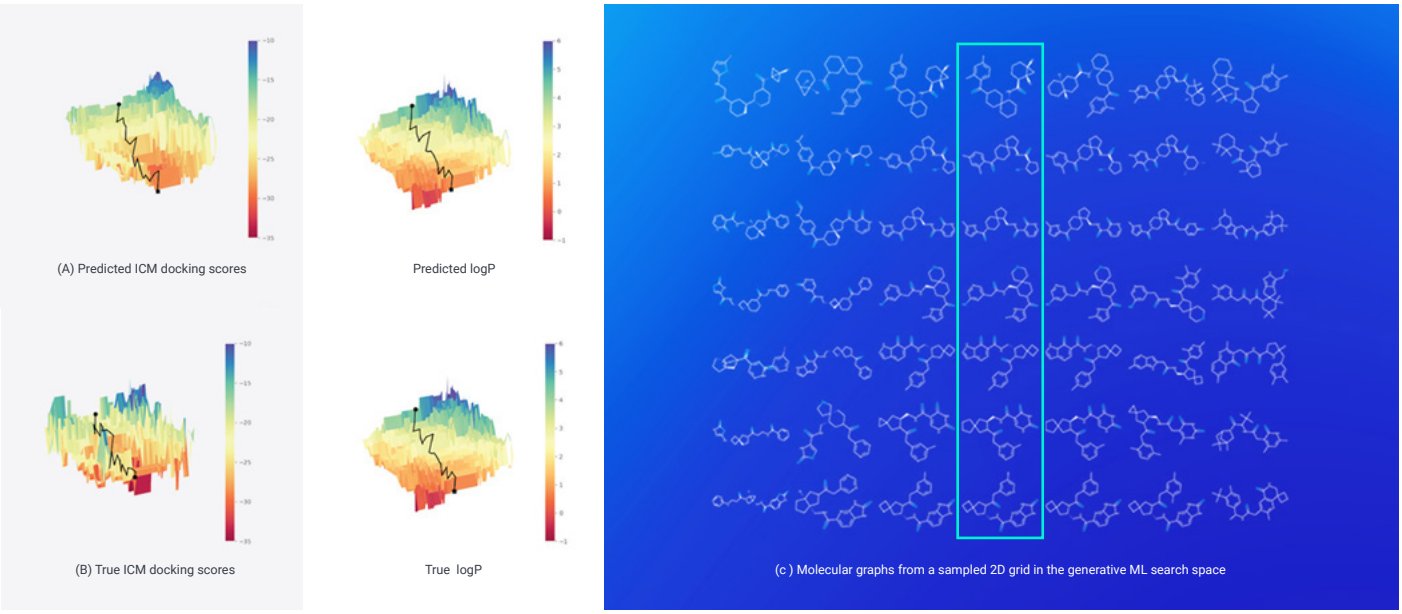


Figure 2 – Multi-property minimization of ICM doscking score and logP in the search space learned by a variational autoencoder. (A) Predicted docking scores and logP, (B) true docking scores and logP, and (C) the corresponding molecular graphs for a 2D slice through the search space. An optimization trajectory through the search space is shown in (A) and (B), and the molecules along the optimization trajectory highlighted wth a blue box in (C).

It is easy to optimize within the search space for many predicted properties jointly, and then map to the associated molecule. Optimization is less likely to become stuck in a local optimum because the search space is much smoother with respect to the properties than the native space of molecular structures. Just as you can find the top of a hill without exhaustively searching its entire surface by repeatedly taking a step in the direction of steepest ascent, optimization within the search space implicitly searches over all molecules that can be represented in the search space. Figure 2 shows

such a search space constructed for ICM docking scores of SARS-CoV-2 3CLpro, a computational proxy for binding affinity that can be evaluated exhaustively over a dense grid of molecules.

Other popular generative ML algorithms include generative adversarial networks (GANs).[13] GANs train a decoder mapping from the search space to molecules, so that molecules decoded from random points in the search space cannot be distinguished from real drug-like molecules by another ML algorithm. GANs were the first generative ML algorithm to produce high-resolution images, but they suffer from mode collapse: the GAN requires that all generated outputs look realistic, but is satisfied even if only a minuscule fraction of possible realistic outputs are ever generated. Moreover, GANs require severe approximations to accommodate discrete domains like molecular graphs, rather than continuous domains like RGB pixel intensities. While GANs have been very popular for both images and molecules, variational autoencoders offer distinct advantages for drug discovery.

The Impact Of Generative Ml On Lead Discovery

Generative ML allows chemical space to be searched more broadly than is possible with conventional techniques. HTS and VHTS are limited to fixed libraries that are small relative to the entirety of chemical space, and are biased toward heavily explored and patented regions. Lead optimization only considers small chemical changes from the hits identified by HTS or VHTS. In contrast, generative ML can search the full extent of make-on-demand libraries with tens of billions of compounds. It can even search over less-explored but still synthesizable regions of chemical space, in which novel chemical matter remains to be discovered, and patents are less dense.

ML can even search over less-explored but still synthesizable regions of chemical space, in which novel chemical matter remains to be discovered, and patents are less dense.

Generative ML can optimize many properties simultaneously within the search space, as shown in Figure 2. Rather than maximizing potency against the primary target alone, as is standard for hit discovery with HTS or VHTS, generative ML can jointly optimize hits for selectivity, ADMET, and physicochemical properties as well. These critical pharmacological properties are often in conflict; for instance, optimization for potency tends to increase molecule size, lipophilicity, and potency for related off-targets. It is difficult for medicinal chemists to actively balance the effect of a chemical modification against a dozen different criteria; generative ML naturally accounts for these trade-offs. As a result of this joint optimization, hits identified by generative ML are expected to be more likely to be true, developable hits. Joint optimization for selectivity tends to eliminate assay interfering, reactive, or aggregating compounds because they are unlikely to be selective. Moreover, because jointly optimized hits from generative ML are already designed for selectivity and ADMET from the start, they should be easier, faster, and less expensive to optimize. Altogether, generative ML promises to reduce the cost and increase the speed of drug discovery.

References

1. Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. Pharmacological reviews, 66(1), 334-395.
2. Ertl, P. (2003). Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. Journal of chemical information and computer sciences, 43(2), 374-380.
3. Bohacek, R. S., McMartin, C., & Guida, W. C. (1996). The art and practice of structure-based drug design: a molecular modeling perspective. Medicinal research reviews, 16(1), 3-50.
4. Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., ... & Tropsha, A. (2014). QSAR modeling: where have you been? Where are you going to?. Journal of medicinal chemistry, 57(12), 4977-5010.
5. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.
6. Gadaleta, D., Mangiatordi, G. F., Catto, M., Carotti, A., & Nicolotti, O. (2016). Applicability domain for QSAR models: Where theory meets reality. International Journal of Quantitative Structure-Property Relationships (IJQSPR), 1(1), 45-63.
7. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., ... & Yu, N. (2021). PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers. arXiv preprint arXiv:2111.12710.
8. Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., & Schmidt, L. (2020, November). Evaluating machine accuracy on imagenet. In International Conference on Machine Learning (pp. 8634-8644). PMLR.
9. Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research, 44(D1), D1045-D1053.
10. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.
11. Cruz-Monteagudo, M., Medina-Franco, J. L., Perez-Castillo, Y., Nicolotti, O., Cordeiro, M. N. D., & Borges, F. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde?. Drug Discovery Today, 19(8), 1069-1080.
12. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

Authors

Dr. Jason Rolfe is CTO of Variational AI, which uses state-of-the-art machine learning in a data-efficient method to rapidly generate novel and diverse compounds that are optimized for multiple properties to avoid the most common causes of drug attrition and increase the probability of clinical success. Variational AI works with leading biopharmaceutical partners and is developing its own internal pipeline. He earned his BS in Mathematics from the Massachusetts Institute of Technology, his PhD in Computation and Neural Systems from the California Institute of Technology, and spent 2 years as a post-doctoral researcher at New York University studying machine learning under Yann LeCun. He has been conducting research in machine learning for more than 15 years, with a focus on deep learning, generative modeling, and cheminformatics.

Dr. Ali Saberali is a machine learning researcher at Variational AI. He earned his PhD in Electrical and Computer Engineering from the University of British Columbia. His research interests are in optimization theory, machine learning, and deep generative modeling.

Mehran Khodabandeh is a machine learning researcher at Variational AI and a PhD student at Simon Fraser University. He earned his MSc from Simon Fraser University in Computer Science.

Corporate Headquarters

577 Great Northern Way, #210
Vancouver, BC V5T 1E1
Canada

variational.ai