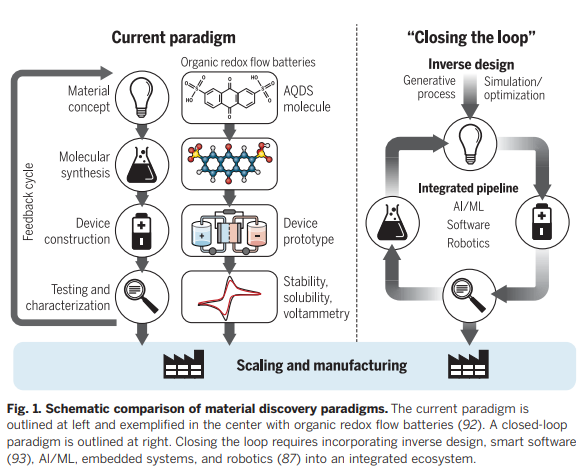
# Inverse molecular design using machine learning: Generative models for matter engineering

* Abstract: - the discovery of new materials can bring enormous societal and technological progress. In this paper researchers review methods for achieving inverse design, which aims to discover tailored materials from the starting field of artificial intelligence.
* In this they mostly examine the approaches to inverse molecular design are being proposed and employed at a rapid pace.
* Introduction: -potential drugs either were prepared by synthesis in a chemical laboratory or were isolated from plants, soil bacteria, or fungus. E.g., up until 2014, 49% of small-molecule cancer drugs were natural products or their derivatives.
* One of the largest collections of molecules, the chemical space project has mapped 1664 billion molecules that contain at most 17 heavy atoms.
* For pharmacologically relevant small molecules, the number of structures is estimated to be on the order of 1060. Therefore, any global strategy for covering this space might seem impossible.
* Simulation offers one way to probing this space without experimentation. The physics and chemistry of these molecules are governed by quantum mechanics, which can be solved via the Schrödinger equation to arrive at their exact properties.
* The size of chemical space is still overwhelming, and smart navigation is required. For this purpose, machine learning (ML), deep learning (DL), and artificial intelligence (AI) have a potential role to play because their computational strategies automatically improve through experience.
* ML techniques are often used for property prediction, seeking to learn a function that maps a molecular material to the property of choice.
* Deep generative models are a special class of DL methods that seek to model the underlying probability distribution of both structure and property and relate them in a nonlinear way. By exploring patterns in massive datasets, these models can distill average and salient features that characterize molecules.
* Inverse design is a component of a more complex materials discovery process.
* The traditional approach takes 15-20 years to discover the product and it takes following steps –
  + Generate a new or improved material concept and simulate its potential suitability;
  + Synthesize the material;
  + Incorporate the material into a device or system; and
  + Characterize and measure the desired properties.
* This cycle generates feedback to repeat, improve, and refine future cycles of discovery.



* The ultimate aim is to concurrently propose, create, and characterize new materials, with each component transmitting and receiving data simultaneously.
* Inverse Design: - Quantum chemical methods reveal properties of a molecular system only after specifying the essential parameters of the constituent atomic nuclei and their three-dimensional (3D) coordinate positions.
* In the inverse design the input is the functionality and the output is the structure.
* Inverse design uses optimization, sampling, and search methods to navigate the manifold of functionality of chemical space.
* One of the earliest efforts in inverse design was the methodology of high-throughput virtual screening (HTVS).
* HTVS starts with an initial library of molecules built on the basis of researchers’ intuition, which narrows down the pool of possible candidate molecules to a tractable range to a thousand to a million.
* Initial candidates are filtered on the basis of focused targets such as ease of synthesis, solubility, toxicity, stability, activity, and selectivity. They also filtered by expert opinions. Successful motifs and substructures are further incorporated in future cycles to further incorporated in future cycles to further optimize functionality.
* HTVS is focused on data-driven discovery, which incorporates automation, time-critical performance, and computational funnels; promising candidates are further processed by more expensive methodologies.
* In organic photovoltaics, molecules have been screened on the basis of frontier orbital energies and photovoltaic conversion efficiency and orbital energies. In organic redox flow batteries and photovoltaic conversion efficiency and orbital energies.
* For organic light-emitting diodes, molecules have been screened for their singlet-triplet gap and photoluminescent emission.
* An optimization approach is preferable to HTSV because it generally visits a smaller number of configurations when exploring the manifold of functionality.
* Within discrete optimization methods, Evolution Strategies (ES) is a popular choice for global optimization and has been used to map chemical space.
* ES involves a structured search that incorporates heuristics and procedures inspired by natural evolution. At each iteration, parameter vectors in population are perturbed and their objective function value evaluated.
* ES has been likened to hill-climbing in high-dimensional space, following the numerical finite difference across parameters that are more successful at optimizing the fitness.
* ES can be quite successful at hard optimization problems, even overcoming state-of-the-art machine learning approaches.
* In other cases, inverse design is realized by incorporating expert knowledge into the optimization procedure, via improved Bayesian sampling with sequential Monte Carlo, invertible system Hamiltonians, deriving analytical gradients of properties with respect to a molecular system, optimizing potential energy surfaces of chemical systems, or discovering design patterns via data-mining techniques.
* Another approach involves generative models stemming from the field of machine learning. The differences between generative and discriminative models are – a discriminative model tries to determine conditional probabilities (*p*(*y*|*x*)): that is, the probability of observing properties *y*, given *x* (a molecular representation).
* By contrast, a generative model attempts to determine a joint probability distribution *p*(*x*, *y*): the probability of observing both the molecular representation and the physical property.
* By conditioning the probability on a molecule (x) or a property (y), we retrieve the notion of direct (*p*(*y*|*x*)) and inverse design (*p*(*x*|*y*)).
* The deep generative models are more challenging to create than direct ML approaches, but DL algorithms and computational strategies have advanced substantially in the last few years, producing astonishing results for generating natural-looking images, constructing high-quality audio waveforms containing speech, generating coherent and structural text, and most recently, designing molecules.
* The researchers focus on three main approaches: variational autoencoders (VAEs), reinforcement learning (RL), and generative adversarial networks (GANs).
* Representation of molecules: - to model molecular systems accurately, the researchers solve the Schrödinger equation (SE) for the molecular electronic.
* From this they obtain properties relating to the energy, geometry, and curvature of the potential energy surface of their system.
* In this model the molecule is represented as a set of nuclear charges and the corresponding Cartesian coordinates of the atomic positions in 3D space.
* At same time the ML algorithm benefit from having representations that expose more easily constraints and properties of the physics of interest, so a 3D representation might not be the most efficient.
* A representation that can span all of chemical space should ideally capture all the symmetries of the SE: permutational, rotational, reflectional, and translational invariance for particles of the same type.
* Current molecular representation fall into three broad categories: discrete, continuous, and weighted graphs.
* For inverse design, a desired property is invertibility- the capability to map back to a molecule structure that can then potentially be synthesized and characterized.
* Following empirical principles of bonding, a molecule is interpreted as an undirected graph where each atom is a node and the bonds are the edges.
* To reduce the complexity, the hydrogen bonds are treated implicitly because they are deduced from standard chemistry valences rules.
* Graphs are normally not uniquely represented, which can be advantageous for data augmentation or disadvantageous when this representation degeneracy introduces noise to a model.
* The Hamiltonians rely only on the known physics and atomic constants of a molecule, the Coulomb matrix representation is based on Coulombic forces between charges of each atom.
* Other representations are better suited to prediction and could be rendered invertible via lookup tables: bag of bonds, amons, fingerprints, electronic density, symmetry functions and chemical environments.
* Generative models for exploring chemical space: - molecular representations are often inputs for deep neural network (DNN) models.
* The DL is a form of representation learning because the DNN architecture is optimized to transform the original data into another representation that is more efficient for a given task such as regression, classification, or generation.
* Each observed datum (molecule) has a corresponding latent representation, often a vector, within a latent variable space that encodes the relevant semantic features of the data.
* The goal of a generation model is to model a data distribution, by training a model on large amounts of data and attempting to generate data like it.
* The loss function encodes the notion of likeness, measuring the differences between two distributions, the empirically observed and the generated.
* The researchers center most of their discussion on deep generative models using SMILES as a representation. For the generation of sequences, recurrent neural networks (RNNs) serve as a common starting point, creating sequences incrementally one step at a time and predicting what comes next.
* Variational autoencoders, reinforcement learning, and adversarial training: - the generative process must be controlled or biased toward desirable qualities.
* By comparison, with GANs and RNNs, the optimization of properties, typically with RL by rewarding or penalizing generative behaviors.
* VAEs give control over the data generation via latent variables. An autoencoder (AE) model includes an encoding and a decoding network.
* The encoder acts as a compression and the decoder as a decompression operation.
* In the act of distilling and expanding information, the AE is expected to learn some of the essential features of the data.
* The AE is sufficient to reproduce the training data, but it can easily learn to memorize the data.
* The VAE achieves better generalizability by constraining the encoding network to generate latent vectors following a probability distribution on the latent space; often the distribution is Gaussian, owing to its accessible numerical and theoretical properties.
* Therefore, a molecule is represented not as a fixed point but as a probability distribution over latent space.
* The most interesting part of a VAE is the latent space. the latent space encodes a geometry; for a given molecule, the researchers’ sample nearly to decode similar molecules. And with increasing distance, they decode increasingly dissimilar molecules.
* Initially, VAEs were proposed for encoding characters of SMILES and then extended to take into account grammar and syntax features, which improve the generation of syntactically valid structures.
* Latent space allows for direct gradient-based optimization of properties, as latent space is a continuous vector space.
* By jointly training the VAE to reproduce molecules and properties, in a semi-supervised fashion, the latent molecular space reorganizes itself so that molecules with similar properties are close to each other.
* Another way of building a generative model is with adversarial training under the GAN framework. In this, the generator competes against a discriminative model; specifically, the generator tries to generate synthetic data from sampling a noise space, whereas the discriminator tries to distinguish data as synthetic or real.
* To bias the generation process with GANs and RNNs, a gradient is needed to guide the optimization of a network toward desired properties.
* To incorporate properties from chemoinformatic tools, simulations, or experimental measures, the researchers need to create a gradient estimator that can backpropagate the generator.
* The field of RL provides several approaches to this problem; among the most prominent are Q-learning and policy gradient.
* RL considers the generator as an agent that must learn how to take actions within an environment or task to maximize some notion of reward.
* The SMILES assigning rewards can only be done once the sequence is completed, to overcome this problem, Monte Carlo Tree Search is often used as it constructs a tree of probabilities and weights, simulating several possible completions for sequences, evaluating their reward, and weighting paths through the tree based on their success or failure at the given task.
* It should be noted that most results of generative models have been used in a pharmaceutical context, optimizing properties relevant to potential drugs such as solubility in water, melting temperature, synthesizability, and presence or absence of certain substructures.
* SMILES represents only the subset of possible molecules; e.g., a syntactically invalid SMILES string might still be a valid molecular structure, but its physics is not encoded by basic valence, but its physics is not encoded by basic valence rules as used in SMILES.
* The introduction of more molecular representations and easy-to-use molecular property predictors will expand the use of generative models in other molecular contexts.
* The VAE framework has been extended to molecular graphs and message passing networks are used to incrementally build graphs.
* There are many challenges remain; it is not yet clear how one can deal practically with approximation methods for the graph isomorphism problem.
* Improved sequence generation models are possible with the ability to read and write to memory. These approaches demonstrate better ability for learning long- and short-term patterns.
* New approaches also lie in inverse RL, geared toward learning a reward or loss function.
* Outlook: - the tools for inverse design, especially those stemming from the field of machine learning, have shown rapid progress in the last several years and have allowed chemical space to be framed into probabilistic data-driven models.
* Generative models produce large numbers of candidate molecules, and the physical realizations of these candidates will require automated high-throughput efforts to validate the generative approach.
* The combination of inverse design tools with active learning approaches such as Bayesian optimization can enable a model that adapts as it explores chemical space, which allows for expanding a model in regions of high uncertainty and enabling the discovery of regions of molecular space with desirable properties as a function of composition.
* Central to machine learning methodologies is the representation of molecules; representations that encode the relevant physics will tend to generalize better.

Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik