# Defining and Exploring Chemical Spaces

* Abstract: - this review provides an overview of some algorithms approaches to defining and exploring chemical spaces that have the potential to operationalize the process of molecular discovery.
* Conceptualizing Chemical Space: - chemical space can be thought of as the set of all possible molecules or materials.
* Here the researchers focus on their discussion on small molecules rather than periodic materials, biomolecules, and polymers, all of which correspond to distinct ‘chemical spaces’.
* The goal is to identify one or more molecules that exhibit a set of desirable properties.
* The two primary considerations one must make are: (i) how to define the space; and (ii) how to explore the space.
* These two aspects are not independent: if we are repurposing FDA-approved drugs, the researchers’ chemical space is narrow enough that an exhaustive screen may be feasible, but if they have no such restriction, theydefin must employ some strategy to select which molecules to test.
* The number of candidate molecules is too large to explore exhaustively, so on often imposes constraints on chemical space depending on the search strategy, the application, and the practical limitations of cost and time.
* This review examines strategies to define and explore chemical spaces with an emphasis on the role of machine learning and synthesizability constraints.
* Defining and Exploring Enumerated Chemical Spaces: - one approach to molecular discovery is to explore a predefined chemical space: an enumerated list of candidate molecules. In this setting, the two stages of (i) defining the space and (ii) exploring the space are entirely decoupled.
* The researchers may think about this problem as an optimization of an objective function f(*x*), where *x* is a molecule belonging to a distance set *X*.
* Defining finite chemical spaces: - careful selection of X can increase the likelihood that it contains a high-performing module while minimizing the number of low-performing compounds.
* The virtual libraries all represent ‘general-purpose’ chemical spaces with broad biological relevance and are therefore applied to many problems related to drug discovery.
* Privileged fragments for drug-like molecules have been identified through retrosynthesis analysis and automatic fragmentation; the molecules produced by recombining these fragments are intended to look more promising than an enumeration based on graph structure alone.
* Graph-theoretical: - this enumeration of molecular structure has been studied for over a century.
* The Chemical Space Project exemplifier modern exhaustive enumeration of all stable organic molecules containing common atom types up to a certain size.
* Reymond and coworkers have enumerated, analyzed, and released the 166.4 billion structures of up to 17 heavy atoms and published numerous visualization and analyses thereof.
* In addition to the benefits of ensuring that X is relevant to the design objective, the predefinition of chemical spaces lets them impose arbitrary constraints on their contents.
* A practical constraint is the ease of experimental validation: that any candidate can be physically acquired for experimental testing.
* Accessibility is the primary motivation for make-on-demand libraries, these libraries are often enumerated by applying a small number of reaction template defining common single-step transformations to all possible combinations of starting materials; recursive enumeration generates molecules accessible through multiple synthesis steps.
* Lyu and colleagues cite an 86% synthesis success rate of 51 compounds selected from 170 million in the Enamine REAL library enumerated from 130 reaction types; WuXi estimates a 60-80% success rate for their 1.7-billion-member collection generated by 30 reaction types.
* The common reaction types exhibit accuracies above 90% on benchmark datasets. These neural network models can be directly used to enumerate possible products or used to predict regio/stereoselectivity patterns.
* Exploring Finite Chemical Spaces: - there are several approaches to identify the top-performing molecules within them. The simplest strategy is to evaluate every candidate molecule.
* But it is not practical to physically test every compound in the ZINC database. More recent studies have since screened over 1 billion enumerated molecules from the same database.
* Active learning is a popular framework that reduce the overall cost through iterative, model-guided optimization.
* It involves selecting subsets of experiments to perform based on predictions from a quantitative structure-property relationship (QSPR) model: a surrogate model (*x*) that codifies an approximation to *f*(*x*).
* While multiple iterations lead to improved surrogate models, a one-iteration approach can still be very effective.
* A novel antibiotic was recently identified from a drug repurposing collection with fewer experiments than an exhaustive screen this way; a similar one-and few-iteration screen was also used to identify kinase inhibitors.
* Both studies used machine learning models as their surrogates: a directed message-passing network and a Gaussian process using compound representation from unsupervised learning, respectively.
* Defining and Exploring Chemical Spaces On the Fly: - using Bayesian optimization, selecting optimal experiments from billions or trillions of molecules requires an equal number of surrogate model predictions; this constrains the size of pre-enumerated chemical spaces that one can consider with a fixed computational budget.
* Genetic Algorithms (GAs): - GAs are model- and derivative- free optimization routines that ‘evolve’ candidate solutions through stochastic mutation and crossover events.
* In an exemplary study, Venkatasubramanian and colleagues define chemistry-informed operations that allow two parent molecules to cross over, two to be merged, one to randomly permute its backbone or side chains, and one to have functional groups inserted, removed, or translocated.
* Subsequent studies refined this strategy for the generation of novel molecular structures by applying mutation operators on molecular graphs or string representations.
* GAs are closely related to fragment-based design, which generates molecular structures piece by piece through addition operations along.
* Deep Generative Models: - these models maintain an implicit definition of chemical space and have shown tremendous promise for molecular design, as reviewed elsewhere.
* When applied to discovery, these approaches explore chemical space by biasing generation to candidates that are high performing. They are usually pretrained on enumerated chemical spaces to learn the basic principles of molecular generation, chemical validity, and what ‘typical’ molecules look like.
* For RL methods where an agent learns a policy to generate molecules, the value of f(x) can be treated as reward to update the agents’ behavior directly.
* Polykovskiy and colleagues used a generative model to propose 300,000 molecules as JAK3 kinase inhibitors, which were filtered to 5000 using docking, clustering, and medchem filters and filtered to 100 using molecular dynamics; finally, one molecule was hand selected and validated.
* In an application of materials discovery, Sumita and colleagues used RL-based generation and density functional theory (DFT) calculations to propose 3200 molecules with targeted maximum absorption wavelengths, of which 86 passed DFT evaluation and six with known synthetic routes were synthesized.
* The chemical spaces that deep generative models explore are defined by the structures they are able to propose as SMILES strings, molecular graphs, or otherwise.
* Autoregressive generation limits the molecules’ sizes but still produces a massive chemical space.
* Defining and Exploring Synthetically Constrained Chemical Spaces O the Fly: - part of why experimental validation is missing from many computer-aided molecular design studies is the expense of physical experiments.
* Molecules sampled from chemical spaces defined on the fly may be challenging, time-consuming, and/or costly to synthesize.
* Here, researchers discuss a category of techniques for synthetically aware definition and exploration of chemical spaces on the fly by incorporating explicit building block and synthetic constraints.
* Synthetically Aware GAs: - synthetic constraints can be incorporated into GAs by restricting mutation operations or by taking advantage of the structured nature of synthetic routes.
* For the latter, a common approach is reaction-based *de novo* design where candidates are generated through expert-encoded reaction templates.
* This approach tends to use model-free algorithms for optimization like GAs to make incremental changes to molecular structures.
* Synthetically Aware Deep Generative Models: - generative models operating on SMILES token or individual atoms are especially prone to generating synthetically challenging structures.
* Synthetically aware models have emerged that integrate reaction-based *de novo* design with machine learning-based generation by reframing molecular optimization as the optimization of synthetic routes.
* Reactants are proposed by a recurrent decoder that selects available starting materials from a discrete list of options.
* ChemBO does not learn a latent space over reactant sets, but instead uses a graph-based reaction predictor to perform a random walk on a synthesis graph by sampling starting materials and simulating reaction outcomes.
* The optimal product according to a surrogate model (x) is selected for full evaluation by *f*(x). this approach is similar to Bayesian optimization within an enumerated chemical space, but the space is continually growing through on-the-fly reaction prediction at each iteration.
* The two other studies formulate chemical space exploration as a Markov decision process – a sequence of actions that correspond to reaction steps in a linear synthesis – and train RL agents to learn a policy that yields optimal products.
* The first, PGFS, selects the other reactant required for biomolecular reactions, while the second, REACTOR, enumerates all possible second reactants and chooses the one with the highest reward.
* The reliability of synthetic pathways comes with the same caveats as make-on-demand libraries; improving models for reaction for reaction prediction will improve their robustness.
* Concluding Remarks: - the discussion has been the desire to efficiently navigate chemical spaces to avoid exhaustive screening.
* The researchers’ also say that they have been slightly biased towards the discovery workflows that ultimately involve experimental validation.
* This approach one should select for chemical space exploration will depend on the nature of the evaluation f(x), the extent to which domain expertise can narrow down the problem-relevant chemical space, and the time/cost budget.
* Closed-Loop Physical Experimentation and Robotic Laboratories: - the integration of computational experimental design algorithms and robotics laboratories – closing the loop – is a promising paradigm for accelerated scientific discovery.
* In the context of molecular discovery, experimental testing requires that proposed candidates can be physically obtained: for manual evaluation, that they be purchasable or synthesizable from commercial starting materials, and for automated evaluation, that they be available or synthesizable from an on-hand chemical inventory given experimental constraints.
* If researchers have the ability to perform only simple one-step chemistries, they can exhaustively enumerate this space and use Bayesian optimization to select from the list of candidates.
* Platforms for automated multistep synthesis with intermediate purification are still in the proof-of-concept phase, but would theoretically have access to a much larger chemical space than what can be practically enumerated.
* Considerations beyond Synthesizability for Molecular Design: - as computer-aided synthesis planning and predicting chemistry tools become increasingly sophisticated, detailed considerations of synthesis time, ease of parallelization, and utilization of common intermediates might be able to be factored into batched molecular design.
* For settings where compounds will be purchased or outsourced, cost-sensitive Bayesian optimization framework can help to quantify tradeoffs between the information that a new experiment will provide and its price.
* Diversity of Chemical Spaces: - as or more important than the size of a chemical space is its diversity and whether it contains molecules that satisfy researchers’ design objective.
* Medicinal chemistry is dominated by a small number of reaction types, but focusing on building-block diversity might let the researchers use simple chemistries more conducive to automation without sacrificing product diversity.
* Diversity-oriented synthesis (DOS) is an orthogonal approach here the reactions themselves introduce a high degree of structural complexity; clever modulation of reaction conditions can lead to dozens of unique products even given simple amine and carboxylic acid building blocks.
* Simplification of Molecular Structures for On-the-Fly Generation: - a limitation of most algorithms for-on-the-fly generation is their ability to handle stereoisomerism.
* SMILES and graph representation are fundamentally unable to distinguish configurational isomers defined by more than tetrahedral chirality and cis/trans bond isomerism.
* Fragment-based methods that explicitly operate in 3D coordinates or generative models designed to propose individual conformers theoretically overcome this limitation, but there have been few (if any) evaluations of on-the-fly molecular generation where the design objective is sensitive to atropiosomerism.
* Simplification of Design Objectives: - computational approximations to physical properties are rarely able to replace physical testing. Benchmarks’ design objectives do not reflect the true complexity of the problem and may not reveal certain failure modes.
* Design objectives tend to focus on optimizing simple heuristics calculated by fragment-contribution approaches or similarity to a target structure; many algorithms already achieve near-perfect results on such tasks.
* The researchers require a new benchmark task that: (i) contain complex design objectives that are less smooth with respect to molecular structure, exhibit more local optima, and require tradeoffs between competing design objectives; (ii) require the generation of individual stereoisomers; (iii) include metrics of sample efficiency; and (iv) have variable experimental costs of candidate molecules.
* Human Involvement in Computer-Aided Chemical Space Exploration: - computational workflows should not be used for their own sake but because they enable discoveries that are otherwise inaccessible to human researchers, whether that is due to the complexity of the optimization task or simply speed or throughput.
* Human filters and manual selection of compounds is almost always an intermediate step between computational predictions and physical experiments and should remain so until we can capture those considerations algorithmically.
* The influence of human subjectivity can be reduced and the process of chemical space exploration can be further operationalized.

Connor W. Coley