Designing Macromolecules using Machine Learning and Simulations

by

Somesh Mohapatra

Submitted to the Department of Materials Science and Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Polymers and Soft Matter

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2022

Author .....
Department of Materials Science and Engineering
April 8, 2022

Certified by .....
Rafael Gómez-Bombarelli
Assistant Professor
Thesis Supervisor

Accepted by .....
Frances M. Ross
Chair, Department Committee on Graduate Studies

Designing Macromolecules using Machine Learning and Simulations

by

Somesh Mohapatra

Submitted to the Department of Materials Science and Engineering
on April 8, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Polymers and Soft Matter

Abstract

The near-infinite number of possible macromolecules, arising from the combinations of monomers, linkages, and their topological arrangement, contributes to the ubiquity and indispensability of macromolecules. However, such chemical diversity hinders the development of general computational approaches that can be applied to macromolecules. The challenges around representing, comparing and learning over macromolecules are manifold. Current representations provide limited coverage of chemical space, and require significant customization to include non-natural monomers and non-linear topologies. Similarity computation methods are limited to biological macromolecules, incorporate evolutionary bias in scoring, and generally do not extend to unnatural monomers or non-linear topologies. Machine learning models are restricted by descriptors with limited representation capacity.

To address these challenges, we developed chemistry-informed representations for the individual monomer unit and the complete macromolecule to capture both the local chemistry and global topology. Chemical similarity computation methods were developed to compare two or more macromolecules, irrespective of monomer chemistry and topology. A wide variety of unsupervised and supervised machine learning methods, selected according to the macromolecule type, data set size, and task, were used to identify patterns in unlabeled data sets, and map macromolecules to properties in labeled data sets, respectively. Using attribution analysis over the pre-trained models, we interpreted the decision-making process of the models. We applied these tools for *de novo* design, virtual screening, and *in silico* optimization of macromolecules, mostly followed by experimental validation of predictions, for applications ranging from peptides and glycans, to electrolytes and thermosets.

Thesis Supervisor: Rafael Gómez-Bombarelli
Title: Assistant Professor

Acknowledgments

In this thesis, a number of people have helped me grow professionally and personally. Their support, in collaborations and otherwise, has immensely contributed to the quality of scientific advancements, resulting in a large number of research publications and conference presentations.

Navigating the PhD would not have been possible without the support of my advisor, Professor Rafael Gómez-Bombarelli. His infectious enthusiasm has enthralled me since the visit weekend. Over the time, Rafa has helped me grow as a researcher and independent thinker. Starting with code debugging sessions to long discussions of how to go about my career, he has been a beacon of light. Whenever I have approached him to do something outside the conventional PhD track, such as an internship or my MBA, he has taken out time to sit down, discuss, and come to an optimal decision.

During this thesis, I have had the opportunity to be guided by Prof Alfredo Alexander-Katz and Prof Bradley Pentelute, as a part of my thesis committee. Taking classes with Alfredo has been one of the most surreal experiences at MIT. I owe a lot of my intuition about polymers to the wonderful lectures and examples, like hydrogels and polymers networks in diapers. Interactions with Brad in the thesis committee meetings and as a collaborator have always been refreshing, enabling me to understand and appreciate both the minute details and the bigger picture of my work.

I owe a lot to professors and students at MIT who have been gracious to share their data, spend time with me to understand the intricacies of chemistry, chemical biology, and biology of the systems, and take a leap of faith in validating the predictions arising from my models. Starting as a first year student in summer of 2018, Carly Schissel, Justin Wolfe and Colin Fadzen (Pentelute lab), shared their data set of cell-penetrating peptides, working on which I produced one of the key works of my PhD. Nina Hartrampf (Pentelute lab) was instrumental in helping me understand the details of the peptide synthesizer. I would like to thank Joseph Brown, Sarah Antilla and Michael Lee (Pentelute lab) for help with peptidomimetic binders; Nicholas Truex

(Pentelute and Irvine labs) for peptide vaccines; Omar Santiago-Reyes and Daria Kim for glycans (Kiessling lab); Yasmeen Alfaraj, Keith Husted and Peyton Shieh for thermosets (Johnson Lab); Qiao Bo for electrolytes (Johnson and Shao-Horn labs); and Nick Yang for organic photoelectronic molecules (Gómez-Bombarelli lab).

A lot of the work would not have been possible without the collaborations and discussions with members of the Gómez-Bombarelli group. A special thanks to Joyce An, an undergraduate intern in the group, who over the past 2 years has helped a lot in the development of the work on non-linear macromolecules. I have enjoyed collaborating with Simon Axelrod, and would like to thank him for his help in both molecular dynamics and generative model projects. Apart from that, I would also like to thank William Hunt Harris, Wujie Wang, James Damewood, Daniel Schwalbe-Koda, Nicholas Layman and others in the group for their support and helpful discussions in and outside group meetings.

During my PhD, I interned at Google, working with Team Sequin. I would like to acknowledge the support of my managers, Lucy Colwell and Suhani Vora, during the internship which helped me gain an appreciation for simpler machine learning models, and the need for development of robust models. Discussions with Jennifer Wei, David Belanger, and Babak Alipanahi, during and after the internship have helped me to remain connected with the latest developments, and align my work accordingly.

I would also like to acknowledge financial support from different fellowships and organizations for my research work - MIT SenseTime Alliance, MIT Jameel Clinic for Health, MIT Takeda Fellowship, and Novo Nordisk.

Pursuing an MBA at Sloan School of Management as a part of the Leaders for Global Operations (LGO) program, in parallel with PhD, could not have been possible without the support of Thomas Roemer, Director, and other staff, including but not limited to Patricia Eames, Anna Voronova, Katja Trizlova, and Ted Equi. I will always remain indebted to them for their help in coordinating the timelines and managing the requirments of both programs at the same time.

In addition to the people I met as a part of research and academics at MIT, I have been associated with a number of student groups - Graduate Student Council,

Graduate Materials Council, Addir Interfaith Dialogue, Tang Hall Residents Association, MIT Machine Intelligence and Manufacturing Operations, and Pi Lambda Phi. I would like to thank members of these groups and people associated with them who have helped me grow as a person over the nearly four years that I have spent here.

Outside my involvements at MIT, I have been fortunate to have great friends in the area and back home in India, whose continuous support through the ups and downs has made this journey enjoyable. First and foremost, I would like to thank Soumya Tripathy, who has been a great friend, a younger brother, and been a constant source of motivation along the thesis. Over morning and evening *chai* sessions, running to the whiteboard and explaining an approach that I have been thinking of for days, has cleared my head. We have taken multiple trips across national parks, biked for an hour or more in the pandemic to eat at a particular restaurant, drove to little-known restaurants in the middle of nowhere, and in the end, started two mildly-popular Instagram channels covering our love for cooking and eating out. Spending Friday evenings, religiously at this point, with Soumya, Kunal Singh and a group of other PhDs, has helped me remain sane through this period. Apart from them, I would like to thank Saswati Majhi, Shivam Gupta, Aarti Dwivedi, and Richard Church, for the numerous calls, meetings and coffee sessions, during the thesis.

Most importantly, I would like to thank my younger brother, Sovesh Mohapatra, who has whole-heartedly supported me all throughout the PhD. I am also grateful for the support and blessings of God, my parents and grandparents.

List of publications and presentations

Peer-reviewed research papers, **co-first author*

- Mohapatra, et al. Chemistry-informed Macromolecule Graph Representation for Similarity Computation, Unsupervised and Supervised Learning. *Machine Learning: Science and Technology* (2022)

- Schissel*, Mohapatra*, et al. Deep learning to design nuclear-targeting abiotic miniproteins. *Nature Chemistry* 13.10 (2021): 992-1000.

- López-Vidal*, Schissel*, Mohapatra, et al. Deep Learning Enables Discovery of a Short Nuclear Targeting Peptide for Efficient Delivery of Antisense Oligomers. *JACS Au* 1.11 (2021): 2009-2020.

- Mohapatra*, et al. Deep learning for prediction and optimization of rapid flow peptide synthesis. *ACS Central Science* 6.12 (2020): 2277-2286.

- Mohapatra, et al. Designing organic photoelectronic molecules with descriptor conditional recurrent neural networks. *Nature Machine Intelligence* 2.12 (2020): 749-752.

- Mohapatra*, et al. Quantitative Mapping of Molecular Substituents to Macroscopic Properties Enables Predictive Design of Oligoethyleneglycol-Based Lithium Electrolytes. *ACS Central Science* 6.7 (2020): 1115-1128.

Pre-prints

- Li*, Zhang*, Mohapatra, et al. Machine Learning Guides Peptide Nucleic Acid Sequence Design. *ChemRxiv* (2021).

Conference presentations

- Mohapatra, et al. Vaccine design using machine learning of degrons. *Neural Information Processing Systems - Learning Meaningful Representations of Life Workshop* (2021).

- Mohapatra*, et al. Interpretable graph representation learning predicts lectin-glycan binding. *Neural Information Processing Systems - Learning Meaningful Representations of Life Workshop* (2021).

- Mohapatra, et al. GLAMOUR: Graph LeArning over MacromOlecUle Representations. *American Chemical Society Fall Meeting* (2021).

- Mohapatra, et al. Graph attribution methods applied to understanding immunogenicity in glycans. *International Conference on Machine Learning - Workshop on Computational Biology* (2021).

- Mohapatra, et al. Chemistry-informed Macromolecule Graph Representation for Similarity Computation and Supervised Learning. *International Conference on Learning Representations - Geometric and Topology Representation Learning Workshop* (2021).

- Mohapatra, et al. Machine learning for advanced affinity selection of peptidomimetic binders. *Neural Information Processing Systems - Learning Meaningful Representations of Life Workshop* (2020).

- Mohapatra, et al. Inferring cell-penetrating peptide design principles using deep learning. *Gordon Research Conference: Chemistry and Biology of Peptides* (2020).

- Mohapatra, et al. Deep Learning for functionality optimization and synthetic accessibility of peptides. *MIT Materials Day* (2019).

- Mohapatra, et al. Inferring cell-penetrating peptide design principles. *Neural Information Processing Systems - Learning Meaningful Representations of Life Workshop* (2019).

- Mohapatra, et al. *Ab initio* non-covalent interaction energies to quantify structure property relationships in lithium-conducting oligomers. *Materials Research Society Fall Meeting* (2019).

In-preparation

- Mohapatra, et al. Shape2Mol: Inverse design of molecules with desired chemistry from shape.
- Mohapatra*, et al. Data-Driven Property Prediction for Accelerated Discovery of Molecular Additives for Robust yet Chemically Deconstructable Thermoset Plastics.
- Mohapatra*, et al. Graph representation learning predicts and interprets lectin-glycan binding.
- Mohapatra*, et al. Machine learning accelerates peptidomimetic binder discovery using affinity selection-mass spectrometry.
- Truex*, Mohapatra*, et al. Vaccine design using machine learning of human degrons.

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Macromolecules - Background and Challenges

1.1 Introduction

Macromolecules are ubiquitous, from constituting what we are made up of to being present in almost everything we use. An individual macromolecule is a result of its monomer composition, connecting linkages, and the spatial arrangement of the monomers and linkages [1]. Individual monomers and linkages are functions of atomic composition, connectivity and stereochemistry, while spatial arrangement of the monomers and linkages dictates the topology. The total number of possible macromolecules grows exponentially as the number and type of monomers, linkages, and arrangement increase. For instance, there are n^{n-2} possible macromolecules, for n unique monomers and just one type of linkage [2]. As a result of such chemical diversity, representing, comparing, and learning over macromolecules with different monomers, bonds, and topologies emerge as critical challenges.

1.1.1 Biomacromolecules

Biological macromolecules are the basis of life, playing a vital role in a wide range of biological processes necessary for survival and growth [3]. They are usually di-

vided into three classes - poly-nucleotides (DNA, RNA), poly-amino acids (proteins), and poly-saccharides (glycans) [4]. Combinations of these main classes of biomacromolecules, commonly referred to as bio-hybrid macromolecules, have been both found in nature and synthesized by humans [5, 6]. Some prominent examples are glycoproteins [7, 8], glycan-RNAs [9, 10], and peptide-nucleic acids [11, 12], .

1.1.2 Synthetic macromolecules

Synthetic macromolecules are man-made macromolecules, with defined chemistry and/or topology. Humans have engineered the composition [13–15] and topology [16] to design structural components [17, 18], sensors [19], shape-memory materials [20], drugs [21], encode messages [22], and much more. Experimental and computational practitioners have explored a vast macromolecule chemical space by varying monomers [23, 24], linkages [25], and topologies – linear [15] and non-linear such as branched [26], star [27], miktoarm [28], and bottle-brush [29].

1.2 Representation

Representing macromolecules in a machine- as well as human-readable form, while trying to encompass enormous chemical diversity, poses a great challenge. Several notations and standards have been developed to address these challenges, both for linear and non-linear macromolecules.

Some of the notable macromolecule representations are strings, hierarchical editing language for macromolecules (HELM) [30], International Union of Pure and Applied Chemistry (IUPAC) international chemical identifier (InChI) [31], CurlySMILES [32] and BigSMILES [33], where SMILES is Simplified Molecular-Input Line-Entry System.

Natural, linear biological macromolecules, such as proteins and DNA/RNA, are usually represented as sequences of one/three-letter monomer codes. This representation is limited by the coding system, and does not have a universally accepted way of encoding new monomers. In a recent attempt, glycans, which are non-linear biologi-

cal macromolecules, were represented as sequences, where groups of monosaccharides were clubbed into 'glycowords' and placed in hierarchical brackets [34].

InChI is a general chemical representation rather than one specifically for macromolecules. This representation encodes the formula, connectivity, isotopes, stereochemistry, and tautomers, in successive layers. As the size of InChI can grow significiantly depending on the size of the molecule, InChIKey was developed as a hash-based representation for indexing and database searching.

HELM, CurlySMILES, and BigSMILES are hierarchical representations inspired by the original SMILES representation [35]. Both HELM and CurlySMILES do not capture the stochastic nature of random macromolecules and require significant customization to be adopted for new classes of macromolecules. BigSMILES is a more general macromolecule representation following the SMILES encoding; however, the limited capability to process using available computational tools restricts its widespread adoption.

Hierarchical fingerprinting is another approach, which follows a hierarchy of atomic (categorical encoding of the presence/absence of contiguous atom-sets), physicochemical (e.g., fraction of rings, molecular surface area), and morphological (e.g., length of side chain, length of main chain) descriptors [36].

In summary, line notations do not always support all topologies, necessitate a fair amount of customization, and have non-canonical variants, while hierarchical fingerprinting is limited in its coverage of chemical space, ability to differentiate stereochemistry, and capturing of long range through-space interactions.

1.3 Similarity computation

Computing the chemical similarity of two or more macromolecules is key to removing redundancies in experiments and leveraging prior knowledge within a class of macromolecules to develop more optimal macromolecules.

For linear biological macromolecules, such as proteins, DNA/RNA and linear glycans, there are several works for computation of sequence similarities [34, 37–39]. Usu-

ally, sequence alignment is done using Smith-Waterman [40] or Needleman-Wunsch [41] algorithms, and scored with substitution matrices, such as BLOSUM62 [42] (for proteins) and GLYSUM [34] (for glycans). These substitution matrices are based on evolutionary statistics, thereby biasing the scoring towards the statistical frequency of a particular monomer's occurrence in the course of evolution rather than chemical similarity. Apart from sequence alignment in linear macromolecules, edit distances [43, 44], linear kernels [45, 46] and deep learning methods [47, 48] have been proposed to compute similarity.

For non-linear macromolecules, alignment of glycans has been explored using q-grams [49], tree matching methods [50–52], and using tree kernels [53]. On the computational front, in recent times, there have been significant advances in computation of similarity using graph edit distances (GED) [54] and graph kernels [55, 56]. Development of software packages, such as graphkernels [57] and GraKeL [58], has provided fast implementations of graph kernels. However, these are general computational methods, and have not been adapted for macromolecules.

Unfortunately, all of the aforementioned methods are limited to biological macromolecules, and do not extend to the general macromolecular chemical space. Moreover, existing tools do not allow incorporation of unnatural monomers and non-linear topologies, except for glycans.

1.4 Machine learning

Machine learning (ML) has enabled a paradigm shift in the field of chemistry, resulting in prediction and optimization of both functionality and synthetic accessibility of molecules for a wide range of applications, ranging from medicine to clean energy [59–61].

ML applications to individual macromolecule classes, such as RNAs and proteins, have been very successful in predicting structure [62, 63] and function [47, 64]. However, these methods typically rely on sequence-based representations that are tailored for linear macromolecules [65]. In a similar vein, PolymerGenome and similar works

have explored using monomer features [66–68] and hierarchical fingerprints to predict different macromolecule properties, like glass transition temperature [36] and dielectric point [69].

However, these methods are mostly limited to linear topologies and a small set of monomers. They are not applicable to the general class of macromolecules, with non-linear structures and diverse monomer and linkage chemistry. Mostly, the lack of an optimal representation limits the development and performance of ML models.

1.5 Problem statement, and thesis overview

In this thesis, we focus on addressing the challenges around representation, similarity computation, and ML of macromolecules by developing methods that cater to a wide range of macromolecules, from sequence-defined macromolecules to networks. These methods are then applied to applications, from the design of mini-proteins for gene therapy delivery to sustainable thermosets to to Li-battery electrolytes. We extended our work to include small molecules, designing organic photoelectronic molecules and drug-like molecules.

The thesis is divided into 5 distinct chapters, other than the Introduction. Chapter 2 introduces the methods, namely, the tools developed or adapted for representation, similarity computation and ML. Chapters 3, 4, and 5, discuss applications of these methods to biomacromolecules, artificial macromolecules, and small molecules, respectively. Chapter 6 concludes the thesis and provides an outlook for the future.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Methods

2.1 Representations

2.1.1 Monomer

Monomers are represented using cheminformatic descriptors and extended connectivity fingerprints (ECFPs). Both descriptors and fingerprints are generated using the open-source Python library RDKit [70].

Physicochemical descriptors quantitatively capture the attributes of the molecule (Figure 2-1A). Some examples are molecular weight, partition coefficient, number of rings and cycles, topological polar surface area, and partial charges. The descriptors, when used to train interpretable supervised ML models, provide an intuitive measure to understand the effect of certain molecular attributes on the final property. Apart from physicochemical descriptors, shape descriptors, such as weighted holistic invariant molecular descriptors (WHIM), are used to represent conformers as fixed-size vectors [71]. However, there are two major limitations in using descriptors - (1) they provide limited information about the molecule due to a selection bias in choosing what descriptors need to be used, and (2) certain descriptors might not be readily available for a lot of molecules, necessitating computationally costly quantum chemical calculations or wet-lab experiments.

Fingerprints capture the connectivity and stereochemistry of the atoms in their

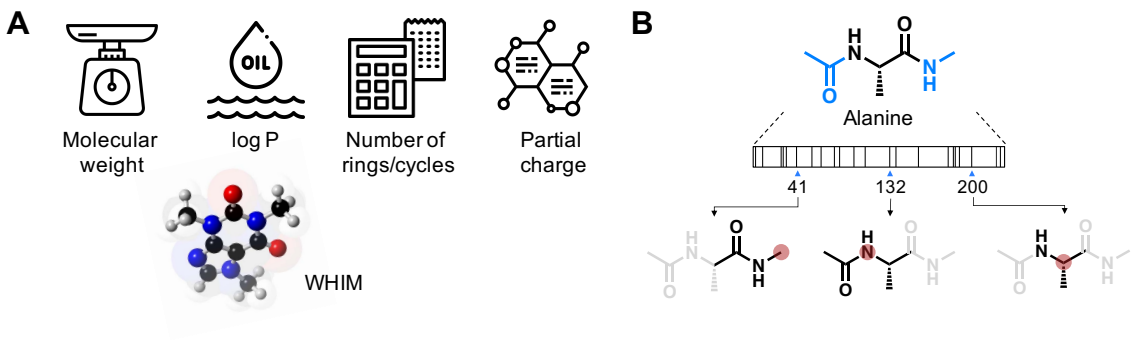


Figure 2-1: **Monomer representations.** A. Selected physicochemical and shape descriptors. B. Circular fingerprint for alanine.

respective neighborhoods in the form of a bit-vector [72] (Figure 2-1B). This vector is obtained by representing the molecule as a 2D graph, where the atoms are nodes and bonds are edges. Each bit in the fingerprint corresponds to a particular sub-graph of the molecule. The value of the bit is 1 or 0, depending on the presence or absence of the specific sub-graph. The fingerprints are unique to a particular molecule, and can be obtained for any new molecule, unlike physicochemical descriptors. For interpretable models, the predicted property can be attributed to sub-graphs or chemical motifs, thereby providing very specific information on the impact of chemical structures on the macroscopic property.

2.1.2 Macromolecule

Macromolecules are represented on the basis of the differentiating group, composition, or sequence/topology-definition (Figure 2-2). Each representation is optimized for the specific learning objective.

Representing the differentiating group is ideal for macromolecules where only a molecular substituent changes while the polymer backbone remains identical for the system. Representing the substituent alone captures the changes differing from the backbone baseline, and avoids redundant representation of the backbone. We used this approach to capture the effect of molecular substituents on conductivity and viscosity of oligoethylene-based Li-battery electrolytes [66].

For polymer networks, the composition and concentration of different monomers

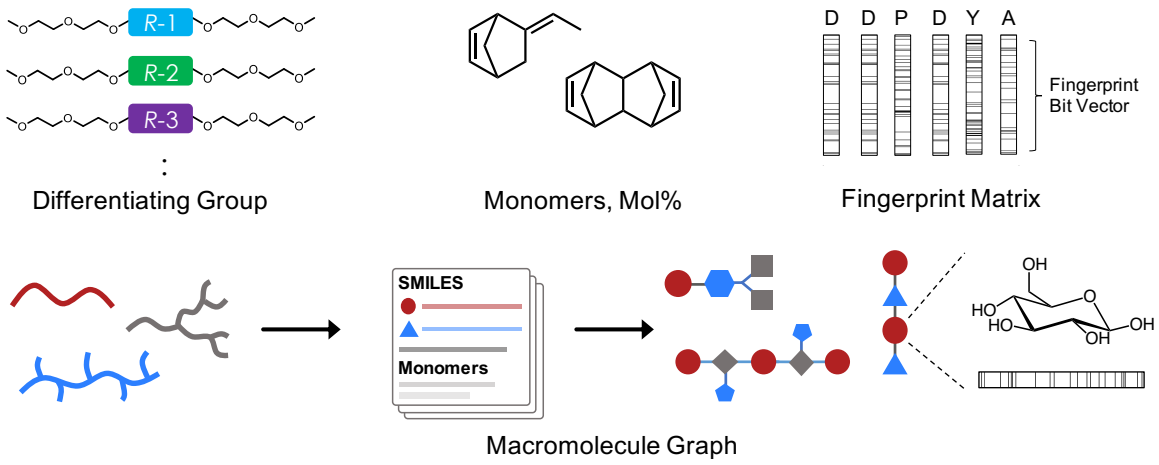


Figure 2-2: **Macromolecule representations.** Differentiating group for molecular substituent in polymer backbones, monomers and their respective concentration for networks, fingerprint matrix for sequence-defined linear macromolecules, and graphs for non-linear macromolecules.

are critical. Thus, representing the monomers using descriptors or fingerprints, and their respective concentration, effectively captures the chemistry of the network.

As the macromolecule gets more defined, such as sequence-defined macromolecules like proteins or DNA/RNA, and sequence/topology-defined macromolecules like glycans, representing the exact position of the monomers and their spatial arrangement becomes essential. For linear macromolecules, a matrix of fingerprints, with the fingerprints arranged in the order the monomers are present in the polymer backbone, captures the local and through-space interactions. Analogously, non-linear macromolecules are well-characterized by macromolecule graphs, where monomers are represented by the nodes and linkages by the edges. The nodes and edges in the macromolecule graph are featurized using fingerprints.

2.2 Similarity computation

2.2.1 Linear macromolecules

Computing similarity for linear macromolecules, represented as sequences of monomer codes, follows a two-step process - sequence alignment and scoring (Figure 2-3A). The

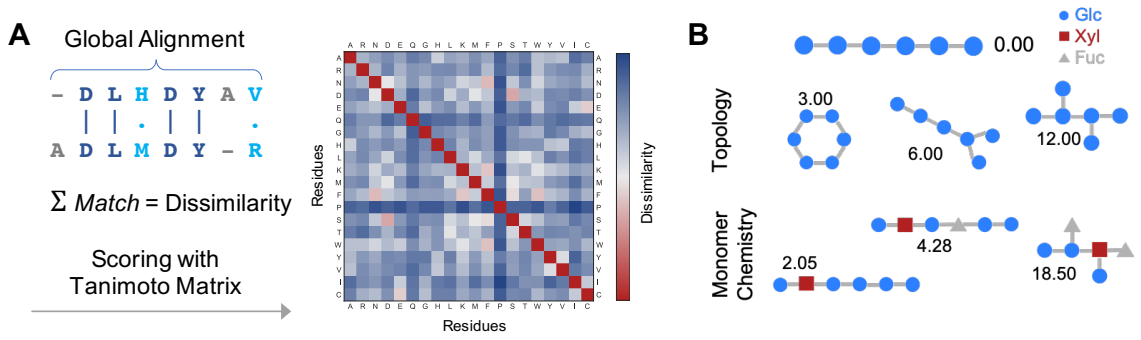


Figure 2-3: **Similarity computation.** A. Global sequence alignment and scoring using Tanimoto matrix for linear macromolecules. B. Graph edit distance for non-linear macromolecules, varying in topology and monomer chemistry.

alignment is solved using a dynamic programming algorithm with $\mathcal{O}(n^2)$ time and space complexity. This approach usually solves for global alignment, which is suitable for relatively shorter sequences like peptides and proteins. Some of the common algorithms are Needleman-Wunsch [41] and Smith-Waterman [40], which have been implemented for alignment of DNA/RNA and protein sequences [37, 38]. The major change in linear sequence alignment is the update to the scoring matrix, where we are using a Tanimoto dissimilarity matrix. The matrix is computed by calculating the Tanimoto dissimilarity between the fingerprint bit-vectors of monomers; thus, self-dissimilarity is 0 and maximum dissimilarity is 1. This approach computes the chemical dissimilarity without any evolutionary bias. Moreover, this is a general approach and can be applied to both natural and non-natural monomers, including those monomers which are not a part of the initial data set.

2.2.2 Non-linear macromolecules

Similarity between two general macromolecules, represented as macromolecule graphs, is computed using graph edit distance (GED) and graph kernel (Figure 2-3B). GED is calculated using A* algorithm, and its modifications use a mixed-integer linear programming approach [54]. The scoring pattern is similar to the one used in scoring between linear macromolecules, based on insertion, deletion and substitution of monomer nodes and linkage edges. We use a Tanimoto dissimilarity matrix to inform

the scores in GED. However, GED is an $\mathcal{NP}$-hard problem, which makes it computationally costly to obtain exact GED as the size of the macromolecules and the data set increases. To scale similarity computation to large data sets, we use graph kernels. Specifically, we use a propagation attribute kernel that propagates the node features along the edges, capturing both local monomer chemistry and global topology.

2.3 Machine learning

In this thesis, we have used a variety of machine learning methods, based on the data set size, data type, and task (Figure 2-4). These methods range from physics-informed models trained on less than 100 unique data points, to language-based models trained on 100,000+ data points. Similarly, the data type, or more specifically the macromolecule type, plays a significant role in the model choice. For non-linear macromolecules, training a graph neural network over macromolecule graphs is more optimal, as it captures the chemistry of the macromolecule in its native state. Ultimately, the objective, such as classification, regression or sampling, dictates the model choice. Often, we benchmark with a variety of other model architectures, and even tasks, such as using a threshold classifier for otherwise regression tasks, to ensure that the initial choice was optimal. In addition, each model undergoes thorough hyperparameter optimization using Bayesian approaches [73, 74].

2.3.1 Unsupervised learning

Unsupervised learning methods have been used to learn the general chemical space using language models [75]. We have used both text completion and translation models to effectively generate novel, yet chemically similar macromolecules and small molecules. The text completion models work by starting with a random seed sequence, and generate the next character, enabling sequence completion following an ontology based on the training data set, thus allowing for sampling of novel linear sequence-defined macromolecules. The translation models enable the mapping of a feature vector, such as a circular fingerprint or a set of properties, to a string-

based representation of the molecule. Both approaches help in taking advantage of large data sets of unlabeled molecules, while the translation approach enables inverse design of molecules with desired properties. We have applied the text completion approach to peptides [64], and the translation approach to design organic photoelectronic molecules [76].

To identify patterns and sub-families in large unlabeled data sets, we have used dimensionality reduction (DR) and clustering approaches. DR helps in the statistical decomposition of the number of input dimensions to a more palatable number of dimensions where the data set can be visualized easily. Linear DR methods, such as principal component analysis (PCA) [77], and non-linear methods like t-stochastic neighbor embeddings (t-SNE) [78], and uniform manifold approximation (UMAP) [79], can be used to transform the data while preserving local and global structures of the data set. Clustering on the low-dimensional data can help in identifying sub-families [80]. The general approach has been used over fingerprint matrices of peptidomimetic binders to identify sub-families, and over glycan similarity vectors to find taxonomic and biologically relevant patterns [81].

2.3.2 Supervised learning

For relatively small data sets, such as those with less than 10 unique molecules, we have used physics-informed models, as they enable learning and extrapolating from known physical relationships. For instance, we successfully modeled the effect of molecular substituents on viscosity and conductivity of oligomer electrolytes using enthalpy-entropy compensation in Arrhenius and Vogel-Tammann-Fulcher (VTF) equations, with a data set of only twelve unique molecules and eight different temperature values [66].

In the low-data regime with 10s to 100s of data points, we have found simpler models, such as ridge [82], decision-tree [83], and gradient-boosting [84] regression, to be more effective adequately capturing the relationship without the risk of overfitting. In addition to that, interpretability using coefficients in linear models, and Gini indices in tree-based models help in attributing the effect of input features on

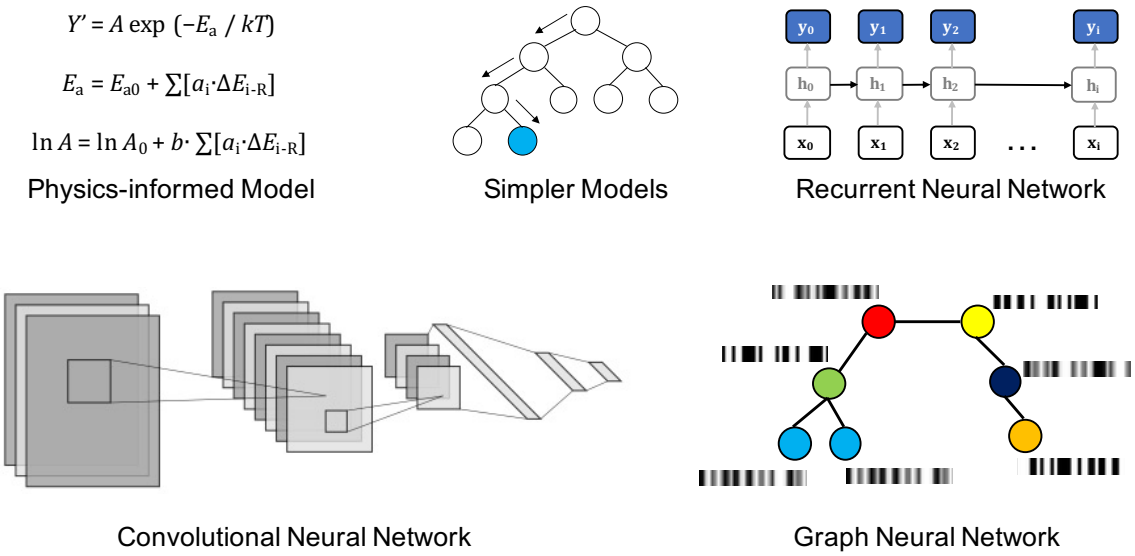


Figure 2-4: **Machine learning.** Different ML model types used for (un)supervised learning in this thesis are shown here.

the predicted property [85, 86].

Convolutional neural networks (CNNs) and graph neural networks (GNNs) are really good at learning over macromolecules where local pair-wise interactions and global through-space interactions are relevant. We have been successful in using CNNs for linear macromolecules, designing nuclear-targeting mini-proteins for gene-therapy delivery [64], as well as using GNNs for non-linear macromolecules, classifying immunogenic and non-immunogenic glycans [81].

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Applications to biomacromolecules

3.1 Cell-penetrating peptides and mini-proteins

Shorter cell-penetrating peptides (CPPs, 5-20 residues) and longer mini-proteins (30-80 residues) have been shown to improve intracellular delivery of drug molecules which otherwise cannot efficiently cross the cell membrane [87–90]. However, the efficacy of CPPs depend on a variety of factors, such as experimental design for *in vitro* and *in vivo* studies, and the identity of the cargo, thereby making it difficult to design general-purpose CPPs [91]. Existing studies in the literature have resulted in inconsistent and sometimes contradictory data sets, owing to the lack of any standard experimental protocols. These erratic and conflicting results have made it harder to formulate sequence-activity relationships, and complicated the development and use of computational methods to develop sequences *de novo* [92–94].

Recent works on quantitative activity prediction of sequence-defined polymers have shown significant promise for the development of computational models to accelerate macromolecule discovery, specifically, for design of anti-microbial peptides and antibody CDR3 loops [95, 96]. For CPPs, there have been attempts to classify peptides as cell-penetrating or not, using binary classifiers [97–100]. However, by the nature of the data set and models, these works are limited to a binary decision, which makes it tough to select peptides for experimental validation in the absence of a quantitative activity. Moreover, these models possess limited accuracy, owing to

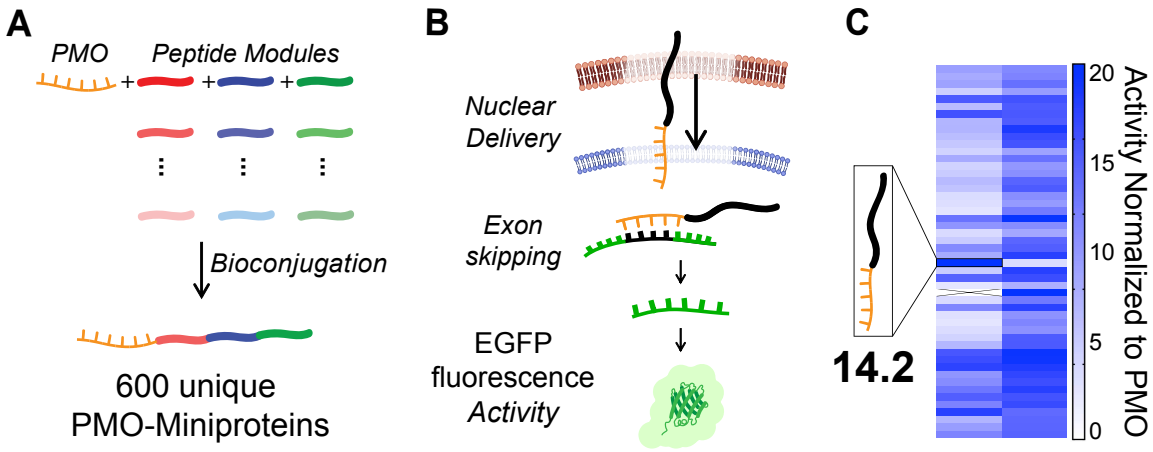


Figure 3-1: **Data set generation.** A. 600 unique mini-proteins conjugated to PMO were synthesized using linear combination of peptide modules. B. Standardized *in vitro* quantitative activity assay tests for nuclear delivery using a quantitative fluorescence readout. C. Members of the modular library exhibit a broad spectrum of activities. Each bit corresponds to a PMO–peptide in the library and its corresponding activity.

challenges arising from the aforementioned inconsistency in the data sources.

In this section, we discuss how we developed an interpretable supervised learning model by leveraging a standardized data set, followed by generation, optimization, and experimental validation of predicted CPPs. We also discuss how we adapted the existing model for generation of shorter CPPs. In addition to being cell-penetrating, both the peptides and mini-proteins are also nuclear-targeting. To be consistent with the convention in the literature, we have continued to use CPP, to refer to the nuclear-targeting peptides and mini-proteins.

3.1.1 Data set

We assembled a standardized, labeled data set by synthesizing and characterizing the nuclear-targeting activity for a library of 600 sequences, conjugated to antisense phosphorodiamidate morpholino oligomer (PMO) (Figure 3-1) [101]. The sequences are linear combinations of modules of short peptides containing diverse structure and function, unnatural residues and non-linear topology in the form of cysteine-linked macrocycles [94]. A high throughput nuclear targeting assay was used to acquire a

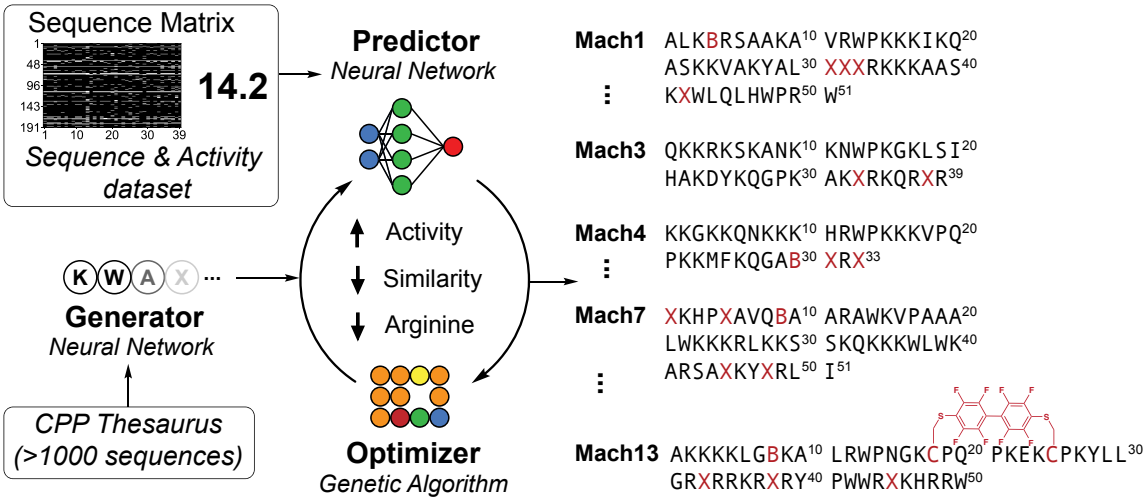


Figure 3-2: **Machine learning framework.** Peptide sequences are represented as fingerprint matrices, and labelled with experimental activity. A CNN model is trained, and used in loop with a genetic algorithm-based optimizer to improve the sequences against an objective function.

quantitative readout of the cell-penetrating activity and nuclear delivery of the PMO. Activity, as quantified by flow cytometry, was reported as mean fluorescence intensity (MFI) relative to PMO alone. Across all the sequences in the data set, the mean MFI was 3, and the maximum was 19.

We also collected a list of CPP sequences from the literature, irrespective of information about penetrating activity [102, 103]. This mostly unlabeled data set was used for training of unsupervised language models.

3.1.2 Machine learning framework

We developed a generator-predictor-optimizer framework to sample mini-proteins, quantitatively predict the intra-cellular delivery efficacy for an unknown sequence, and optimize sampled sequences against an objective function, respectively (Figure 3-2).

Generator. The generator was modeled on a next-character predicting language model, where the training is performed using fixed-size inputs and single-character outputs. For example, the input could be 'MATERIA' and the output 'L'. In this

manner, the model learns the sequence ontology, specifically, the general order of the amino acids in CPPs. When used for sampling new sequences using fixed-size seeds, the model produces novel, CPP-like peptide sequences.

A recurrent neural network - nested long short-term memory (RNN-Nested LSTM) model architecture was used to train the generator [104]. LSTM models are good at remembering long-range interactions, which translates to being able to capture non-local interactions between amino acids from the primary structure [105]. Additionally, the nested architecture helps in attending to both local and non-local interactions, as compared to other simpler RNN architectures, such as vanilla-LSTM or bidirectional-LSTM [104]. The model was trained using a list of known CPPs, both from the library we assembled and existing sequences in the literature [102]. Since it is a language model which needs to learn the CPP grammar, we were able to expand our training data set to include CPPs without quantitative labels.

For training, the characters were featurized as one-hot encodings, where the set of n characters can be assumed as n categories, and each character represented as an n-sized vector of 0s, with a single 1 at a particular index. To sample new sequences using the generator, we used short seed sequences, randomly sampled from the training data set. We noted that the resulting sequences sampled using the generator had low string similarity, calculated using Jaro-Winkler text distance, with the sequences in the training data set. With this approach, we fulfilled one of the key objectives of generating novel sequences [106].

Predictor. Using the labeled data set, we trained a supervised machine learning model, to model the experimental activity from the sequences (Figure 3-3). We used a convolutional neural network (CNN) model architecture for the supervised model. The CNN model works by aggregating the local features into combined representations, essentially approximating the contribution of multiple local features in a single representation. In a sequence scenario, the process starts by capturing pairwise interactions, ultimately leading to the through-space interactions.

The training data set consisted of a curated list of sequences from the assembled

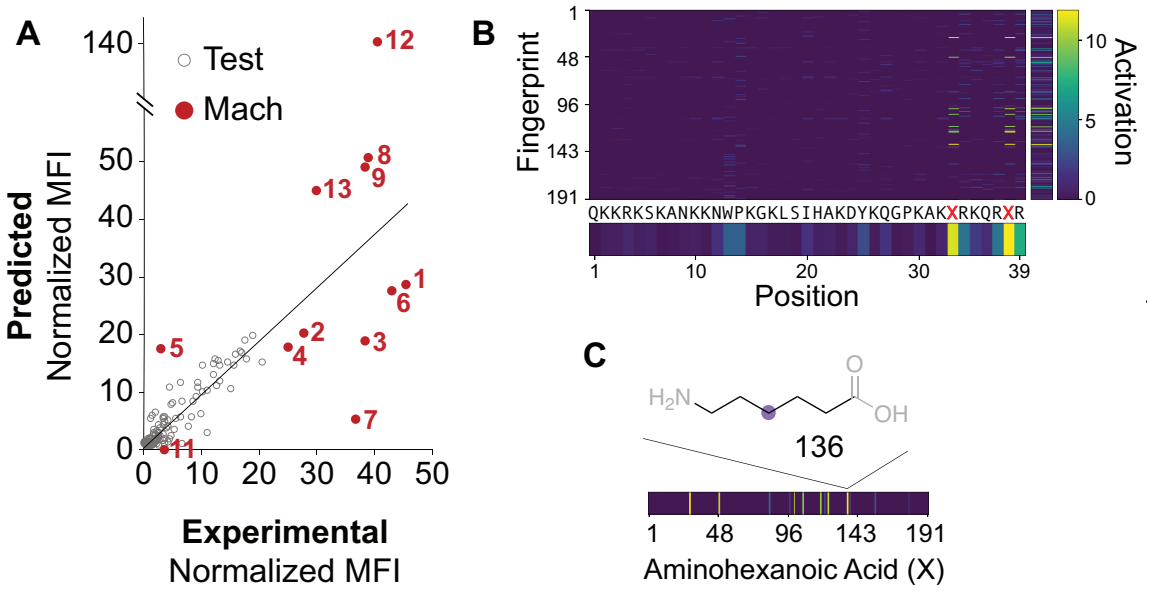


Figure 3-3: **Model performance and attribution.** A. Parity plot showing the predicted and experimental activities for sequences in the test data set, and experimentally validated or Mach sequences. B. Positive activation gradient map showing the activated fingerprint bits for Mach 3, with the averaged activation values across all fingerprint indices and residues shown on the right and bottom of the feature activation map. Amino acids corresponding to the most activated bits, X or aminohexanoic acid, have been marked in red. C. Activation gradient map for aminohexanoic acid, and substructure corresponding to the most activated fingerprint bit.

library, and the individual peptide sequence modules. Sequences were represented as matrices of extended connectivity fingerprints. We randomly split the data set into 80:20 for training and validation of the model.

The model performed well within the range of the MFI values in the training data set, with a root mean squared error (RMSE) for the validation data set being 0.4 of the standard deviation of the training data. We benchmarked these models against a variety of model architectures, classification tasks, and one-hot sequence representations. We noted that almost all the models were limited by the range of the training data, and only the CNN model trained over fingerprints could extrapolate in the codomain space. However, the extrapolation came along with increased noise for the predicted values. We attempted to address the statistical noise by model ensembling, and were able to average out the extreme predicted values.

Optimizer. The directed evolution inspired optimizer completed the synthesize-learn-design loop. A genetic algorithm (GA) based strategy using random mutations was used to optimize the sequences against an objective function. The mutations involved single-point mutations, such as insertion, deletion, or swapping of residues in the seed sequence, and motif mutations or hybridization, involving the replacement of a random-sized motif with another same or different sized motif sampled from the training data set. We performed 1000 iterations for every seed sequence.

The objective function involved maximization of the activity predicted using the pre-trained model, minimization of arginine content as well as length and similarity to the library, all while retaining water solubility -

$$GA_{score} = \frac{1}{2} MFI_{pred} - \frac{1}{2}\left(\frac{1}{2} R_{count} + \frac{1}{5} L - \frac{1}{10} NC + S\right), \tag{3.1}$$

where GA_{score} is the score used to compare the seed and mutated sequences, with the sequence corresponding to the higher value being used as the seed for the next iteration; MFI_{pred} is the activity predicted using the supervised model; R_{count} is the number of arginine residues in the sequence; L is the sequence length; NC is the isoelectric point estimate for the sequence based on the amino acid composition [107]; and S is the mean Jaro-Winkler similarity of the sequence compared to the training data set.

The GA was used to obtain sequences with a range of activities - ranging from negligible improvement of using the CPP, as compared to naked PMO, or approximately 0-fold activity, to 3x the maximum activity observed in the training data set. This approach was taken to assess the robustness of the model in predicting activity of the sequences across the range of training data, apart from its ability to extrapolate. Moreover, the optimization for the lower activity sequences was to check if the model could generate negative control.

3.1.3 Interpreting the predictor model

To interpret the decision-making process of the model, specifically, to attribute the predicted activity to specific residues and molecular features in the input representation, we used gradient class activation maps (GradCAM) [108]. This method has been widely used for attribution analysis for image classification. Briefly, this approach assumes the model as a mathematical function, and obtains the gradients corresponding to each input feature. These gradients are multiplied with the respective features. The values greater than zero are noted to have a positive influence on the predicted result, and those less than zero have a negative influence.

Here, we adapted this framework for regression, generating fingerprint bit-wise positive and negative activation values, corresponding to the effect of the respective feature on the predicted activity. For Mach3, one of the most active CPPs, we observed that the C-terminus aminohexanoic acid (Ahx) residues were most positively activated compared to the rest of the residues. Further, the fingerprint bit for the alkyl chain in the Ahx residue was the most activated substructure.

The level of detail on the effect of individual residues and substructures was unprecedented using both computational and experimental tools. Prior to this framework, arduous experimental methods, such as alanine scanning of the entire peptide sequence, where individual residues are replaced by alanine, were used to evaluate the change in the activity and for attribution to individual residues. The savings, in terms of time and resources, accelerated the development of new hypotheses and directed wet-lab experiments accordingly. In one such experiment, we investigated whether the attribution feature is useful towards post hoc mutations to Mach miniproteins, and found a substantial boost in activity when mutating Ahx (6-carbon chain) to aminoundecanoic acid (11-carbon chain) in Mach3.

To probe the model further, we used this visualization approach to analyze different sequences 3-4. We randomly selected five sequences of different lengths, seeded them in the predictor–optimizer loop to maximize the activity contingent on other design constraints and visualized the activation for the best predictions. In line with

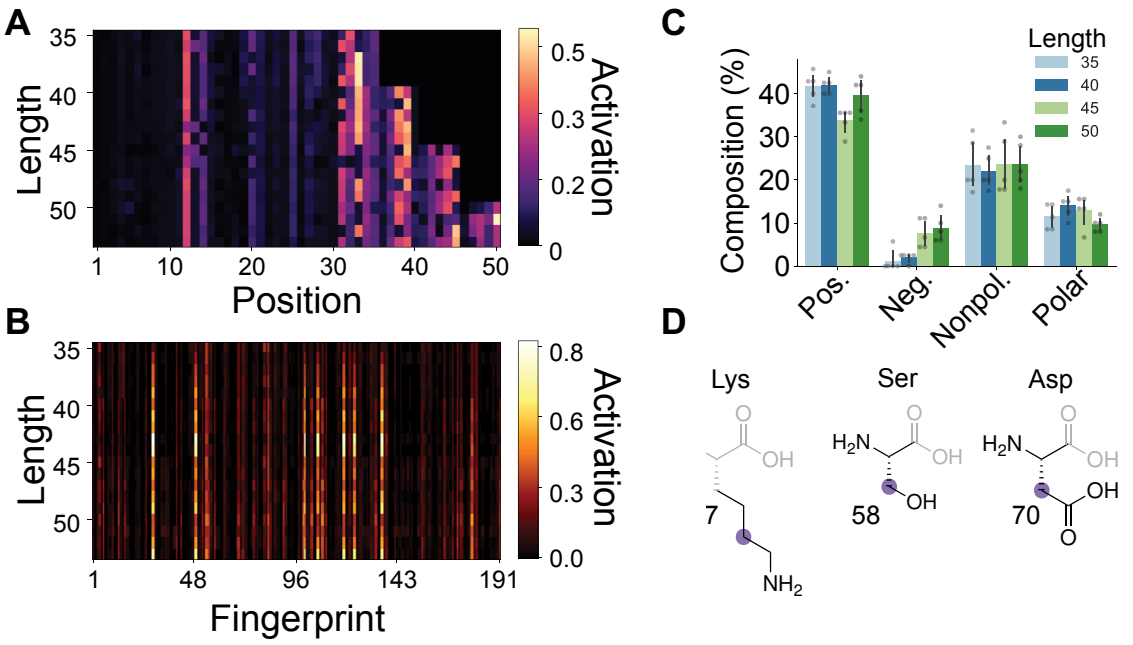


Figure 3-4: **Inferring design principles using attribution.** Feature activation maps, averaged across the A. fingeprint indices and B. residues, shown for the optimized sequences for lengths, 35, 40, 45 and 50. C. The percentage composition of positively-charged, negatively-charged, non-polar, and polar amino acids shown for the sequences. D. Activated substructures for Lys, Ser and Asp are shown. The blue dot indicates the node and the black bond lines indicate the sub-graph in the 2D graph of the amino acid, with the rest of the structure in grey. The number noted alongside each substructure refers to the fingerprint bit index in the extended connectivity fingerprint.

our previous observation for Mach3, we noted a higher activation for residues at the C-terminus. It may be noted that the C-terminus is the free end of the peptide, with the PMO conjugated to the N-terminus. General understanding of CPPs suggests that highly active peptides constitute a good number of positively-charged amino acids to penetrate the negatively charged cell membrane [109]. We also observed that the percentage of positively charged amino acids remained on the higher side. Additionally, the general composition of charged and hydrophobic amino acids remained unchanged across different sequence lengths. At the substructure level, specific fingerprint bits were noted to be activated, irrespective of the sequence length, such as the side chains of lysine, serine and aspartic acid.

3.1.4 *In vitro* validation of nuclear-targeting mini-proteins

We synthesized and characterized twelve sequences from the list of sequences optimized using the model framework. The sequences were selected to capture a wide range of lengths, predicted activities, and arginine contents. Along with sequences predicted to have a high activity, we selected sequences with a lower activity, as negative control, to check the robustness of the supervised model. We used the same assay, as was done for the library generation, to characterize these sequences conjugated to the PMO cargo. All but one of the sequences that were predicted to have activity greater than 20-fold surpassed the best performing sequence in the generated library with a 19-fold experimental activity. Despite the limited amount of data available for training, and the extrapolation outside the range of activities in the training data, we were able to reasonably predict the trend of predicted and experimental activities for the new sequences.

The model was able to capture sequence-activity relationships, beyond correlations of activities with physicochemical properties. In the library used for training the model, longer sequences, and sequences with high arginine content resulted in higher activity. Using the model, we were able to design sequences with approximately 30 residues and very low arginine content, and with activity much greater than the maximum of the training data set. These results demonstrate how machine learning approaches outdo heuristic-driven physicochemical property-based optimization strategies.

3.1.5 Design of shorter CPPs

As an extension to our work on mini-proteins, we intended to design shorter CPPs with less than 20 residues [110]. These shorter peptides are preferable to longer mini-proteins, owing to their low molecular weight and ease of synthesis. However, we have observed a positive correlation between sequence length and activity of the CPP [101]. In addition, the majority of the CPPs in the original training data set were longer than 30 residues, indicating a biased data set with limited information about

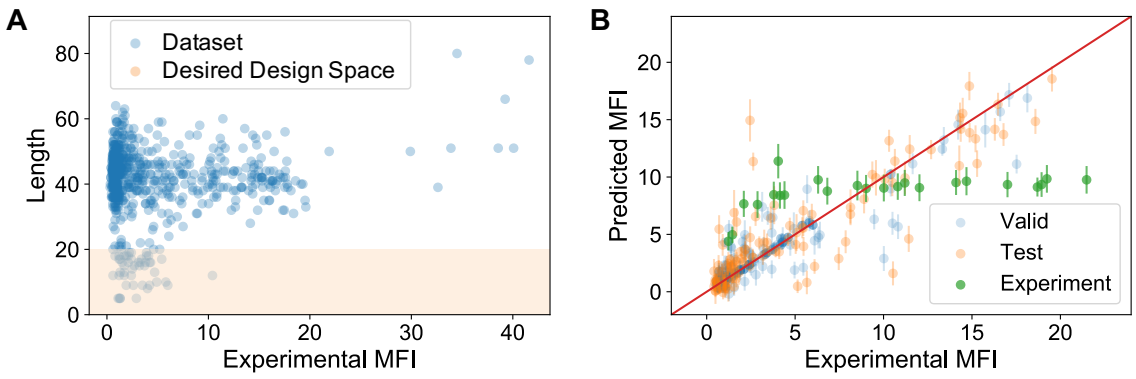


Figure 3-5: **Designing shorter CPPs.** A. 601 out of 653 sequences in the training data set are longer than 20 residues (scatter plot). The desired design space of sequences with 20 or less residues is denoted in orange. B. Parity plot for the 10x data augmented CNN model ensemble is shown, with mean fluorescence intensity (MFI) for sequences in the validation (blue), test (orange), and experiment (green) data sets. The points indicate the mean value of the predicted MFI from all of the models in the ensemble, and the error bars denote the standard deviation of prediction.

sequence-activity relationships in the desired design space. Thus, designing a highly active CPP, with the added constraint of length posed a greater challenge.

To address the length imbalance in the data set, we used data augmentation. The data set used in the training the model in Schissel et al. [64] had 640 sequences, with 92% of the sequences having more than 20 residues, and the mean sequence length being 41. For this work, along with the original 640 sequences, we included the 13 sequences published in Schissel et al. [64]. We replicated the 8% shorter sequences nine times, bringing the total number of each sequence in the data set to ten. The augmentation allowed us to train the model over a desired length space using the shorter sequences, and show the model characteristics of highly active sequences using the longer sequences in the data set.

With this augmentation strategy in place, we were able to see only a slight improvement predictions using the CNN model. The improvement can be noted from comparing experimental activities in the held-out test data set and predicted CPPs validated *in vitro* 3-5, where the R^2 score is 0.148 and the Pearson's correlation is 0.512. Benchmarking against simpler models, such as gradient boosting and multilayer perceptrons, we observed that the simpler models had a better prediction of

mean activity but had significantly large ensemble variance, rendering the mean values unreliable.

In a retrospective active learning experiment, we added sequences generated as a part of the study to train new models. Out of a total of 19 sequences, we added 1, 4, 7, 10, 13, and 16 random sequences in the training data set, and tested the new model on the remaining sequences. We observed that the performance on the test data set improved in subsequent generations of the model. This improvement indicates that active learning with new sequences can help in sequential optimization of the model to predict sequences in desired design space.

3.2 Advanced affinity selection of peptide binders

Many early-stage drug discovery efforts focus on the identification of compounds that potently bind a biomolecular target [111, 112]. Often such compounds are identified by mining large libraries, such as high-throughput screening libraries or libraries generated through combinatorial approaches [113]. A major bottleneck of these techniques, however, is the routine identification of nonspecific binders—compounds that exhibit little or no binding preference for the target of interest relative to other targets during library interrogation [114, 115]. These nonspecific binders can additionally appear as 'hits' across multiple assays, including pan assay interference compounds [114, 116]. This complication requires the activity of each putative binder to be experimentally validated, a time-consuming and labor-intensive process [112]. While some filters exist to give confidence to the identification of functional binders, biophysical evaluation is generally required to truly evaluate each 'hit' [116]. At the most detrimental, these nonspecific putative binders can derail entire screening efforts [115, 117].

Peptides continue to garner further therapeutic interest due to their potential to disrupt clinically-relevant protein-protein interactions (PPIs) [118–120]. Their synthetic accessibility [121], amenability to chemical tailoring [122, 123], and potential for cell penetration [64, 124] have made them increasingly popular starting points in

drug discovery. New peptide binders are typically discovered from libraries accessed via either genetically-encoded or chemically-accessed approaches, such as phage display [125–127], mRNA display [128, 129], one-bead-one-compound (OBOC) screening [130], or affinity selection-mass spectrometry (AS-MS) [131, 132]. Chemically-accessed peptide libraries in particular offer the advantage of easy access to expanded chemical space, enabled by direct incorporation of 'non-canonical' amino acids and amenability to a variety of macrocyclization strategies. However, inherent to all of these techniques is the concomitant discovery of specific and nonspecific binders, despite a number of approaches aimed at addressing this challenge.

Machine learning has presented a paradigm shift to the drug discovery field, accelerating efforts in hit generation by virtual screening [133, 134]. For peptides specifically, there are multiple instances where machine learning models trained over data from generated libraries or public databases have predicted peptides that experimentally demonstrate higher activity [134–136]. Accurate predictions of PPIs have been demonstrated using three-dimensional structures of target protein and peptides [137, 138]. However, these approaches are limited because they require three-dimensional structures of peptides and target proteins for prediction and do not capture protein dynamics or allostery. There is another class of models that predict the binding affinity or probability of binding for target-specific neoantigens using sequence information [139–141]. These models represent amino acids as individual alphabets using one-hot encodings, or with physicochemical descriptors such as charge, molecular weight, etc. However, because these models trained on specific binders alone, they cannot be used to identify nonspecific binders and non-binders. Furthermore, without atomistic representation of the chemical structure in the model, the diverse chemical space encompassed by noncanonical amino acids cannot be effectively included or compared.

Here, we report an unsupervised-supervised machine learning model approach based on topological representation of amino acids to advance drug discovery using affinity selection-mass spectrometry. We used data obtained from affinity selections of canonical L-peptide libraries against IgG 12ca5 used in an recent optimization of AS-

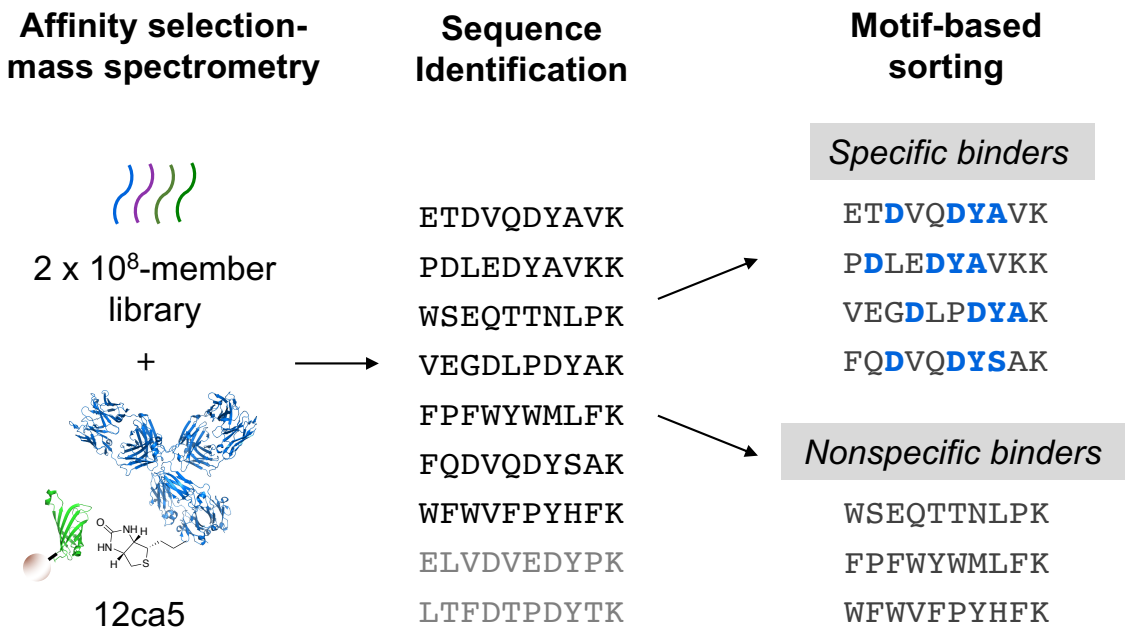


Figure 3-6: **Data set generation.** Specific binders to anti-hemagglutinin monoclonal antibody clone 12ca5 are identified by affinity selection-mass spectrometry (AS-MS) and distinguished from nonspecific binders based on the known characteristic 12ca5-specific binding motifs. Unlabeled sequences in the sequence identification step are colored in grey.

MS [132]. Unsupervised learning readily distinguishes specific and nonspecific binders on the basis of chemical similarity. Supervised learning classifies unknown sequences into specific binders, nonspecific binders, and non-binders. Predicted labels of 12ca5-specific, nonspecific, and non-binders were then validated in AS-MS experiment to reveal the performance accuracy of the supervised learning classifier.

3.2.1 Data set

Affinity selection data was gathered using a recently optimized workflow, described in Quartararo et al. [132] (Figure 3-6). First, the biotinylated target protein (mouse anti-hemagglutinin monoclonal antibody, clone 12ca5) was immobilized onto streptavidin magnetic beads. After washing with blocking buffer, the immobilized target protein was incubated with a 2×10^8 L-peptide library composed of 18 canonical amino acids. Cysteine was excluded to allow only linear peptide architecture and isoleucine was

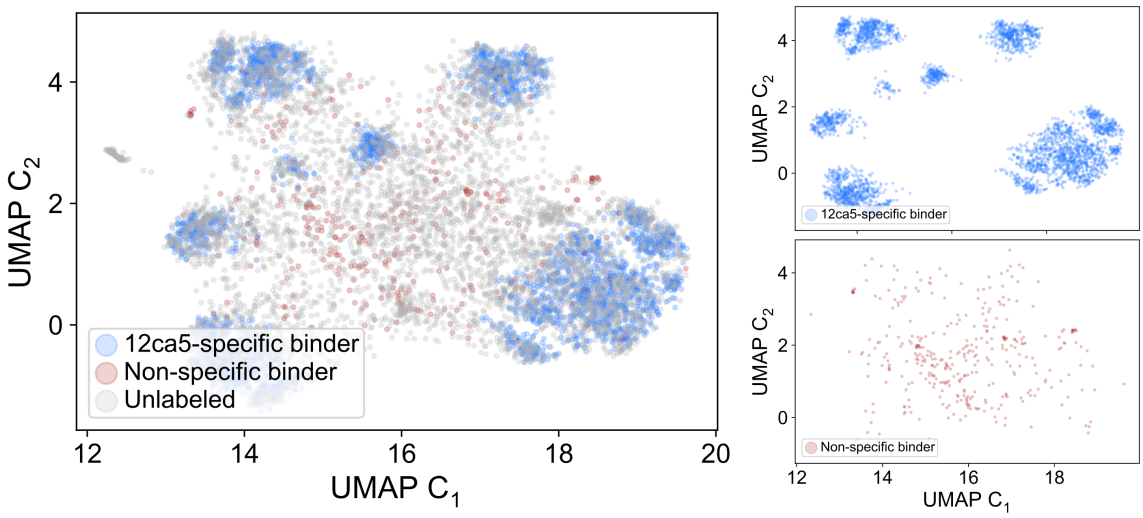


Figure 3-7: **Unsupervised learning over binders.** UMAP over matrices of residue fingerprints differentiates peptide sequences into 12ca5-specific binder clusters, spread of nonspecific binders, and continuum of unlabeled sequences.

excluded as it cannot be distinguished from leucine without isotopic labeling in mass spectrometry. After incubation and washing, 12ca5-bound peptides were eluted and subjected to de novo peptide sequencing via nanoliquid chromatography Orbitrap mass spectrometry [142]. All identified peptides that match the synthetic library design are considered binding peptides (specific and nonspecific), whereas non-binding peptides are removed during the washing steps of the selection. Sequences which did not match the motif were marked as unlabeled. Affinity selection was completed in triplicate. 12ca5 was used for initial AS-MS optimization because it is known to bind linear peptides with the motif of D–DYA/S. Given this known binding sequence, the binding peptides observed by AS-MS were then sorted as 12ca5-specific or nonspecific binders.

3.2.2 Unsupervised learning

Using fingerprint matrices for peptide sequences, we were able to differentiate between 12ca5-specific and nonspecific binders, and identify sub-families within 12ca5-specific binders (Figure 3-7). The fingerprint matrix for each sequence was was flattened into a vector, and decomposed into two components using the uniform manifold approx-

Figure 3-8: **Interactive plot for binder visualization.** A Jupyter notebook-based visualization was created to (de-)select binder groups, check sequences, and even new sequences, such as those marked as Trial_1, and see where they lay.

imation (UMAP) method. The unlabeled sequences formed the continuum between islands of 12ca5-specific binder sub-families in the two-component space. In addition to the fingerprint matrix representation, we also performed unsupervised learning over similarity vector representation, comparing the similarity of each peptide with rest of the library, as described in section 2.2.1. Dimensionality reduction of the similarity vectors was done using multi-dimensional scaling [143].

We benchmarked the UMAP dimensionality reduction against t-stochastic neighbor embeddings (t-SNE), and performed hyperparameter optimization for both approaches. Specifically, n neighbors were optimized for UMAP, and perplexity and number of iterations were optimized for t-SNE. We observed that t-SNE was able to separate the sub-families globally, but there was sparsity observed at a local level. However, UMAP was able to differentiate at both local and global scales. This observation is in line with results reported in literature, comparing t-SNE and UMAP for other dimensionality reduction tasks [79, 81, 144].

To make the pipeline accessible to experimentalists, we created a Jupyter notebook interface to handle data loading, dimensionality reduction, and an interactive plot to

down-select by different labels and identify sequences by hovering over the specific points in the two-component space (Figure 3-8).

3.2.3 Supervised learning classifier

The supervised learning was done in two steps: first, one-class classification and generation of negative data set (unrecorded sequences which might have been washed away during the AS-MS experiment), and second, multi-class classification (Figure 3-9). From AS-MS and motif-based sorting, we obtained a data set of binding sequences (12ca5-specifc and nonspecific binders), while all non-binding sequences were washed away. Thus, there was no explicit information available about non-binding sequences. Therefore, we obtained a data set of non-binding sequences with a sequence generator–one-class classifier loop. We trained an isolation forest one-class classifier using similarity vectors to learn the chemical similarity of binding sequences. A generator was set up to sample sequences of same length as binding sequences. The generator was constrained to use the library of 18 amino acids and have a C-terminal Lys, the same as the experimental peptide library. A similarity vector for each sequence was calculated against the library of binding sequences. This similarity vector was used as an input for one-class classification task. Non-binding sequences were generated to create a balanced number of sequences for the multi-class classification (n = 2704).

We trained a CNN model for classification of specific, nonspecific, and non-binders. The CNN model was trained using bit-vector matrices of sequences against respective class labels. When evaluated against 20% of randomly held out data set during the training, the model had a categorical cross entropy loss of 0.002 and 99% accuracy. 99% accuracy was seen when the model was evaluated against another 20% of the data set which was not used in the training of the model.

The model was able to classify nonspecific binders for 12ca5. We hypothesize that nonspecific binders are common across multiple selections of similar format (e.g., magnetic Dynabeads). However, each protein could have its set of nonspecific binders, indicating our model will improve with additional affinity selections against other

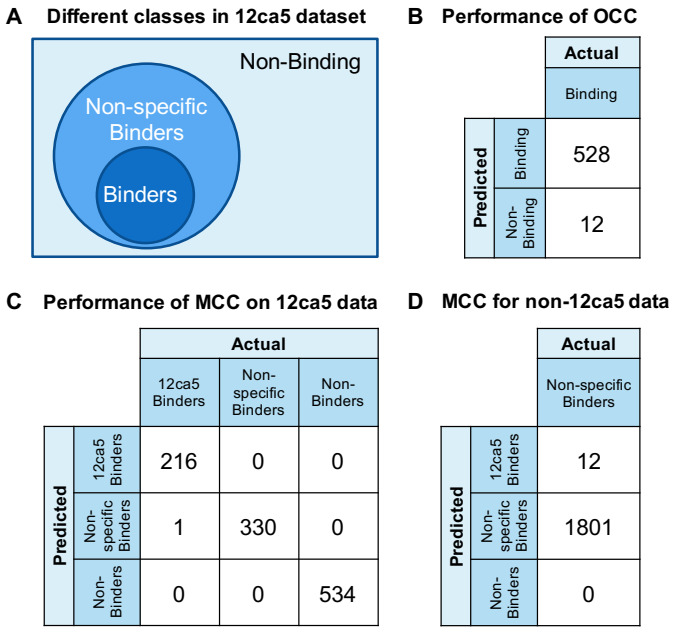


Figure 3-9: **Supervised learning of binders.** A. Different peptide classes are represented in the Venn diagram. There are two resulting classification tasks: i) one-class classification between binding (defined as both 12ca5-specific and nonspecific binders) and non-binding sequences (estimated to be 108 sequences), and ii) Multi-class classification of 12ca5-spcific binders, nonspecific binders, and non-binding sequences. B. Performance of one-class classifier against test sequences demonstrated 98% accuracy. C. Performance of convolutional neural network (CNN) classification model against 20% of test sequences demonstrated 99% accuracy. D. Classification of nonspecific binders from HLA A1101 selection using the model trained on 12ca5 data successfully classifies 1801 sequences as nonspecific and 12 sequences as 12ca5-specific binders.

targets. We used our model trained with 12ca5-binder data to classify nonspecific binders identified in HLA A1101 screening. 1801 out of 1813 sequences (99% accuracy) were classified correctly as nonspecific binders, despite the model having never been trained on the specific data set.

3.2.4 Experimental validation of peptide sequences identified using the supervised model

Using the supervised multi-class classifier model, we generated peptide sequences *in silico* with predicted labels for experimental AS-MS validation that were 'nonobvious' to classify by eye (Figure 3-10). Specifically, peptide sequences were generated (n =

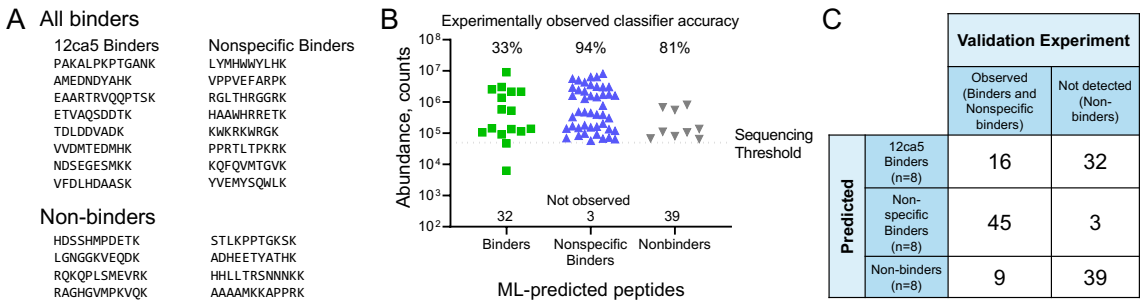


Figure 3-10: **Machine learning-predicted sequences demonstrate high accuracy in experimental validation to classify nonspecific binders.** A. Peptide sequences were predicted for each classification in silico (n = 8 each), synthesized, and purified for validation in affinity selection. All sequences were chosen to be "nonobvious" in classification. Specifically, i) 12ca5-specific binders do not contain the D–DYA binding motif, and ii) both the nonspecific binders and nonbinders span a range of hydrophobicity. B. Upon affinity selection, the abundance of each peptide by mass spectrometry revealed nonspecific binding peptides were accurately predicted (94% accuracy), similar to nonbinding peptides (88% accuracy). Predicted 12ca5-specific binders that do not contain the D–DYA binding motif showed some binding (33% of population), meaning the machine learning model was able to discover new sequences outside the known D–DYA/S binding motif through chemical similarity. C. An outcome matrix explicitly shows the number of times each predicted sequence was observed (meaning it is a binder or non-specific binder) or not detected (non-binder) in affinity selection.

8) for each label considered of 12ca5-specific, nonspecific, and non-binders. 'Nonobvious' nonspecific binders and non-binders were created by varying the predicted peptides over a range of hydrophobicity, which can contribute significantly in the nonspecific binding of peptides to proteins and surfaces [145]. Moreover, 'nonobvious' 12ca5-specific binders were generated by predicting binding sequences that did not contain the D–DYA/S sequence. This requirement simultaneously pushes the model to its limit as well as potentially explores the sequence space of the motif, similar to an alanine scan. The 12ca5-specific sequences predicted demonstrate high chemical similarity to the D–DYA/S sequence, but include frameshifts, deletions, and substitutions.

After completing the AS-MS experiment with the ML-predicted peptides, the abundance of each peptide demonstrated that nonspecific binding peptides were accurately predicted (94% accuracy), along with nonbinding peptides (88% accuracy).

Within our currently optimized methods [132, 142], peptides will be sequenced if they are above an observed abundance of 5×10^4 counts. If the machine learning predictions is accurate, 12ca5-specific or nonspecific binders should be observed at or above this sequencing threshold, while non-binders are washed away in the selection.

The machine learning model generally showed high accuracy for the nonspecific and non-binding peptides. In our experiment, 12ca5-specific binders predicted by machine learning only showed 33% accuracy for binding. However, this can be rationalized by our initial requirement that these predicted peptides should not contain the D–DYA/S binding motif. This result means that 12ca5 binding to peptides is highly sequence specific to the D–DYA/S binding motif.

3.3 Dynamic time warping analysis of affinity selection - mass spectrometry data

High-throughput library screening methods, as discussed in Section 3.2, can identify compounds that bind to the target [111–113]. The pre-processing of the raw data undertaken in these methods often result in significant loss of information, sometimes up to 99% of the compounds [146]. Such a huge data loss represents the loss of valuable information which could have been leveraged better. Affinity selection - mass spectrometry (AS-MS) method used for identification of peptide binders that inhibit protein-protein interactions are also a victim of this phenomenon.

In a typical AS-MS experiment, the peptide library is screened against a control and another target protein. The proteins are conjugated to streptavidin biotin magnetic beads [147]. These magnetic bead bound proteins are incubated with the peptide libraries, usually consisting of greater than 1 million peptides [122, 132]. Using an external magnet, the bound peptide-protein complexes are held to the wall of the container, and thus separated from the unbound peptides which are washed away. After purification, these peptides are then fed into a mass spectrometer that obtains spectra as the peptides elute, based on their mass, followed by a second mass

spectrometry of automatically identified spectra of interest, giving the process the name - tandem mass spectrometry, or MS/MS. The second MS is usually performed if peptide-like features are observed in the primary MS.

Raw data from the MS/MS experiment is fed into a sequencing algorithm, such as PEAKS [148, 149]. The mass spectra is processed on the basis of the intensities of different mass/charge ratios (m/z), resulting in a list of potential bound peptides. Further processing is done to filter incorrect sequence assignments, noise and side products [142]. Less than 100 peptides, out of which a lot of the peptides are usually non-specific, are identified from a starting library of more than 1 million. The non-specific peptides are filtered by comparing against the peptides present in both target and control sets, thereby indicating no specific preference for a particular protein. In most cases, less than 10 peptides are found to be viable hits, or that is not even the case, for an unknown and hard-to-bind target protein.

The loss of information in the processing steps necessitates the utilization of the raw data using a method that amplifies the signal-to-noise ratio, and helps in identification of more hits, as compared to contemporary approaches.

In signal processing research, multi-variate time series data, similar to the AS-MS data, have been analyzed using methods like dynamic time warping (DTW) [150–155]. DTW is a method applicable to almost any set of time series data and was developed specifically to combat the problem of time variation [156]. For instance, if we have two series which expected to have a similar structure but we want some deterministic way to match patterns in one series to similar patterns in the other, the naive way is to simply match identical time indices (i.e. match each of the first observations to each other, then match the second observations, then the third, etc). This approach runs in to problems if the two series do not have the same number of observations, are shifted in time, or occur at different rates. DTW addresses all of these problems and determines the optimal way to match the time indices of each series in a largely brute-force approach under some reasonable requirements. In general, DTW can be thought of as finding the best way to compress and stretch time in one series to make it look like another so that different durations, time shifts, and speeds can all be

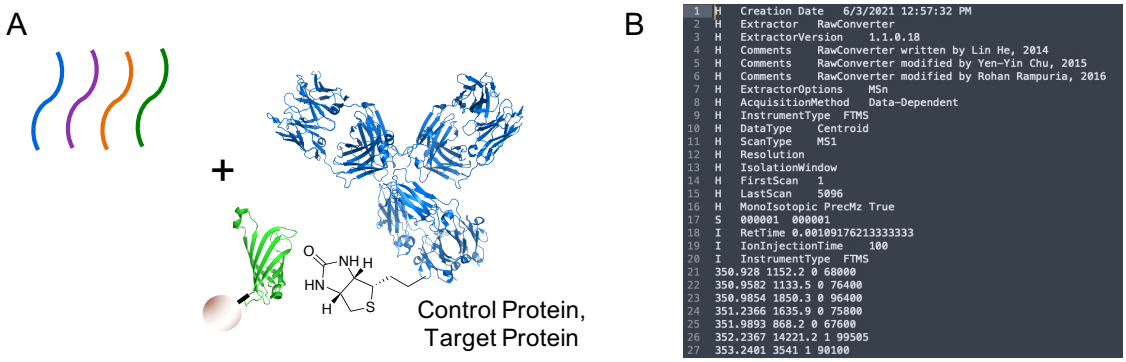


Figure 3-11: **AS-MS process and output file.** A. Schematic of the AS-MS process showing the peptide library, protein conjugated to streptavidin biotin magnetic bead. B. Representative image of MS1 file.

accounted for.

In this work, we formulated AS-MS spectra as a multi-variate time series, a collection of scans over time where each scan contains a collection of m/z values and corresponding intensities. In our work we started with two sets of data, a control protein (12ca5) and a target protein (SARS-Cov2-RBD) consisting of triplicate spectra each. The task was to determine what structures, specifically what intensity spikes, were present in the target group and not in the control group. The challenge is that AS-MS inherently involves small variations in time and the specific structures, short bursts of intensity, are relatively brief occurrences. In order to overcome this time variation, we used DTW to align the series for a single target protein together, amplified the signal by adding the individual spectra, and subtracted the aligned control spectra from the target to identify the unique intensities, thereby identifying the target-specific peptides.

3.3.1 Data set and pre-processing

Raw MS-1 spectra obtained from selection against 12ca5 and RBD proteins were used as-is (Figure 3-11). The raw data is captured using 2 distinct file systems - MS1 and mzXML. The mZXML files document the m/z scan at distinct time periods, irrespective of presence of eluted material. In contrast, MS1 files are a subset of the

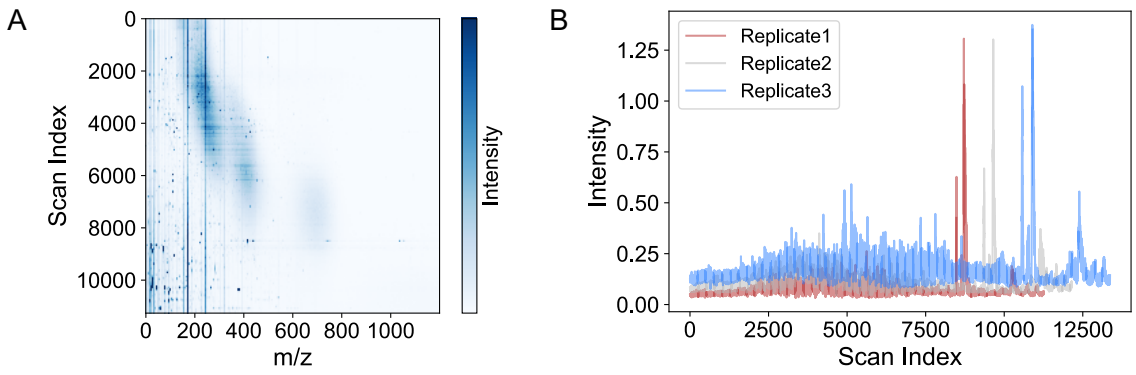
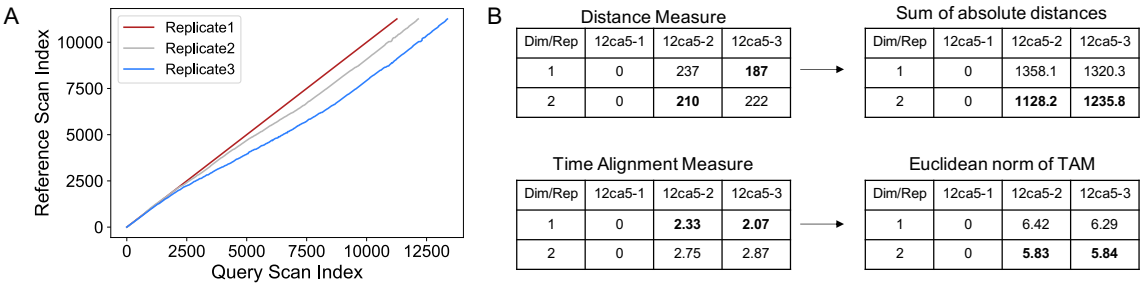


Figure 3-12: **Visualization of MS1 spectra.** A. MS1 spectra of peptides against 12ca5 control protein is visualized for all scan numbers shown on the y-axis, and 200-1400 m/z values, with the x-axis shifted by 200. The coloration shows the intensity of the peak at a particular m/z value for a specific scan. B. 1D spectra for the 3 replicates of 12ca5 affinity selection. The range of scan numbers and peak positions indicate how different individual runs can be. The 1D spectrum is obtained by summing over all m/z intensities at a particular scan index.

mzXML files after filtering out scans which did not include any peptides, containing about 10,000 scans, as compared to 30,000 scans in the mzXML file. In our analysis, we used the smaller MS1 file, instead of the complete mzXML files, for the analysis.

In an effort to impose a consistent structure across all replicates, we sorted the observed m/z values into bins depending on their decimal value, essentially, rounding each m/z value to a given number of decimal places and possibly cutting out values which are too low or too high (Figure 3-12). If two m/z values rounded to the same number, we kept only one copy of that number but their corresponding intensities were summed. The upper limit for number of decimal places in our data is four, but due to memory constraints we have usually rounded to zero decimal places - the nearest integer. In benchmarks for different decimal places, we noted that the rounding off did not result in any significant change in the alignment or downstream analysis. For each replicate this binned data was ultimately stored as a 2-dimensional array.



Dim/Rep	12ca5-1	12ca5-2	12ca5-3
1	0	237	**187**
2	0	**210**	222

Dim/Rep	12ca5-1	12ca5-2	12ca5-3
1	0	1358.1	1320.3
2	0	**1128.2**	**1235.8**

Dim/Rep	12ca5-1	12ca5-2	12ca5-3
1	0	**2.33**	**2.07**
2	0	2.75	2.87

Dim/Rep	12ca5-1	12ca5-2	12ca5-3
1	0	6.42	6.29
2	0	**5.83**	**5.84**

Figure 3-13: **DTW Alignment and metrics.** A. Warping path of each replicate for all the 12ca5 replicates, when warped using DTW. Replicate 1 is selected as the reference, and all others are aligned to it. The red line is the self-alignment of replicate 1, hence a straight line. B. Different metrics of alignment - distance and time alignment measures of one-to-one alignment measures, followed by sum of absolute distances.

3.3.2 Alignment of different replicates

We used multi-variate time series alignment, as implemented in the dtw-package [157], to align the different replicates. To understand the difference in the amount of data required for alignment, we aligned replicates of reduced 1D spectrum and complete 2D spectrum. The 1D spectrum was obtained by summing over all corresponding intensities of a particular scan, resulting in an intensity versus scan plot. In contrast, the 2D spectrum had individual m/z values and intensities for each scan. We arbitrarily assigned the first replicate of each protein as the reference spectrum, and aligned the rest of the replicates to it. By adding all the aligned spectra together, we obtained the amplified spectrum. However, before the analysis, it is important to note how well the alignment is.

Most often DTW is used to determine how similar two series are, giving a single distance value rather than the particular time index matches. This DTW distance has been used to measure similarity in speech recognition, signal processing, signature validation, and many other areas where small variations in speed and pauses should not largely affect similarity [156]. Although the DTW distance is usually used, the first step in its calculation is finding the optimal time alignment or warping path. That alignment can be used directly to identify and match the patterns in each series.

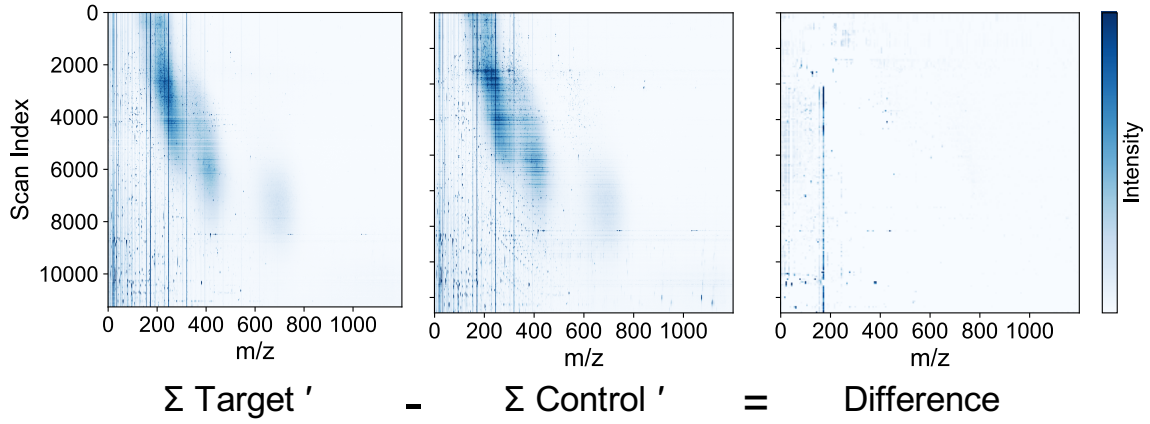


Figure 3-14: **Difference of aligned spectra** The aligned sum of replicates for the control protein is subtracted from the target protein to obtain the difference spectra.

In our work we primarily used the alignment/warping path or time alignment measure (TAM), and also calculated the DTW distance as one measure of how good the alignment is (Figure 3-13). In addition to one-to-one alignments, fitness of final product of these alignments in terms of the actual data is essential. Since each alignment can produce a warping path, multiple aligned data sets, and a final difference data set and each data set can either be a 1D or 2D time series, we used many different measures of fitness to understand the alignment.

3.3.3 Difference of aligned spectra

DTW distance is the distance between reference and aligned query [152]. This is the distance (Euclidean by default) between two series after they've been aligned through time warping. By the definition of DTW, this is the minimum (Euclidean) distance between the series across all allowed time warpings. It is a non-negative number with no upper bound except that given by the data. This approach can be applied to either the 1D data set or the 2D data sets although the DTW distance between 1D chromatograms should not be compared to the DTW distance between 2D arrays since the way Euclidean distance is calculated in each of these cases makes it an unfair and misleading comparison. This is why we apply 1D alignments to the 2D data set or vice versa; if we do that so that the dimension of each data set is the same and we

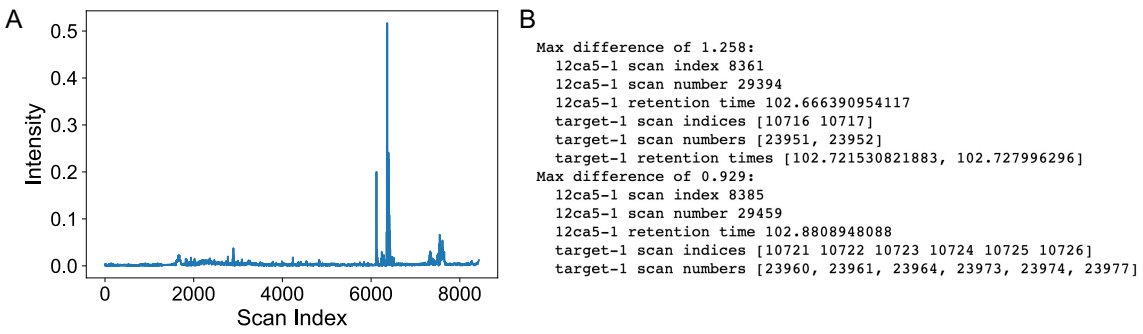


Figure 3-15: **Target-specific scan indices.** A. 1D chromatogram showing the difference of intensities at different scan indices. B. Representative image of scan indices in the Jupyter notebook.

recalculate DTW distance then the comparisons are fair.

TAM for a given warping path is a measure of how much two series have to be warped in order to attain the DTW distance [158]. It is a number between 0 and 3 where 0 indicates that no warping was needed (probably the two series are scaled versions of each other) while 3 indicates that the two series were never in phase (the warping path contains no diagonal steps). This measure can be calculated using only the warping path of an alignment and does not have the same dimensionality problem as the DTW distance. Further, converting a 1D-aligned chromatogram to a 1D-aligned matrix will not change the warping path. So we do often directly compare TAM values of 1D alignments to TAM values of 2D alignments.

We noted that the results of the alignment varied across different metrics. In one-to-one alignments, the most consistent measure was TAM, and we could see that alignment of 1D spectra was better than the 2D spectra. However, visual inspection of warping paths overlaid with the original time series indicated that 1D alignment was not robust enough and could be corrupted by the noise in the series. The measures were consistent for the summed spectra, with 2D alignment being preferable to 1D across all metrics. This approach showed that the usage of the complete spectra, or a multi-variate time series alignment, worked better than using a single summary series.

To obtain the target-specific scans, we aligned the amplified spectrum of the con-

trol protein to the target, and subtracted the control from the target (Figure 3-14). This subtracted spectrum shows, for each scan, the specific m/z values that were observed only in the target data set, and the m/z values which had a larger intensity in the target compared to the control. Summing over the intensity values across the m/z values, we obtained the information about specific scans and thus the spectra unique to the target protein.

The aforementioned pipeline enables identification of target-specific peptide sequences from raw spectra obtained from AS-MS (Figure 3-15). To help the experimentalists expedite their AS-MS analysis, we built a Jupyter notebook interface to handle the raw data, process the replicates using DTW alignment, and display target-specific scan. In another experiment, the m/z values obtained from the analysis of scan indices could be used to improve the parameters in the tandem mass spectrometry step, where specific masses could be concentrated on for thorough analysis.

3.4 Vaccine design using machine learning of human degrons

Personalized molecular vaccines have enabled a paradigm shift in therapeutic strategies for cancer and other pathogenic diseases by catering to individual patients [159, 160]. In general, these vaccines are developed by a well-established protocol (Figure 3-16) [161–164]. First, a sample of the tumor is acquired by biopsy, or other invasive procedure, followed by sequencing for pathogen characterization. Comparing the sequencing data with a healthy sample, the mutations are identified, and the immunogenic antigen is selected. Accordingly, vaccine sequences are developed and formulated for administration. Once injected, these sequences activate the innate immune system which in turn lead to the desired therapeutic effect at the tumor site.

In the above process, neutralizing antibodies are readily produced by the extracellular antigens, however, intracellular antigen processing and presentation remains critical for developing memory in T-cells, a part of the immune system [165–168].

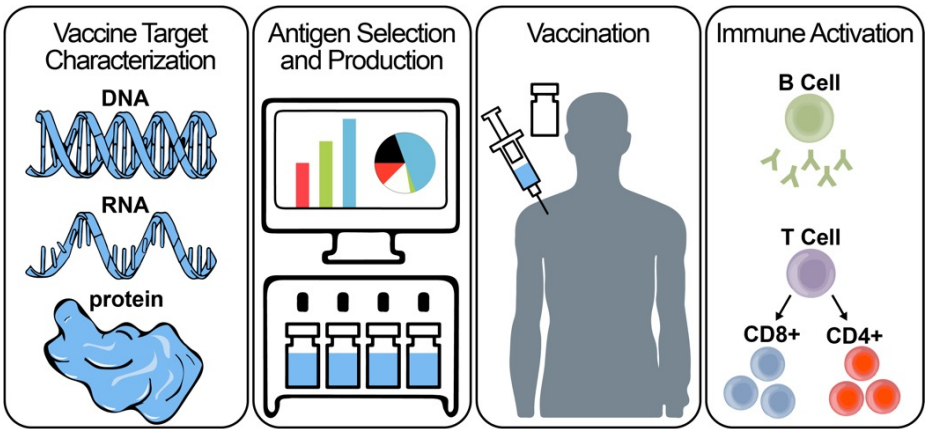


Figure 3-16: **Vaccine design.** Development of a molecular vaccine through characterization of DNA, RNA, or protein sequences; selection and production of immunogenic antigens; administration of vaccine components; and immune priming of antigen-specific T and B cells.

The intraceullular processing follows ubiquitinylation and proteasomal degradation in the cellular cytoplasm, N-terminal trimming in the endoplasmic reticulum, and loading on to the class I human leukocyte antigen (HLA) molecules. Antigens bound to the HLA localize at the cell surface, and presented to the T-cells. At this point, the T-cells recognize the antigens and are activated.

The two-step process necessitates that the vaccine peptides are processed well by the proteasomal degradation system for immune activation, and high HLA-binding affinity [169, 170]. Although a few optimization approaches for proteasomal processing have been developed, their successful implementation remains a significant challenge [171, 172]. For HLA-binding affinity prediction, robust predictors based on bioinformatics and machine learning are available and routinely used by the immunotherapy community [173–175]. Hence, improved vaccine design strategies, that combine both antigen processing and presentation are needed for improved immune recognition.

Harnessing the degron pathways in the ubiquitin proteasome degradation system may offer a favorable approach to facilitate antigen processing (Figure 3-17). In this system, ubiquitin ligases catalyze and regulate the ligation of peptides to the ubiquitin protein, and further proteasomal degradation [176]. In a study, as early as

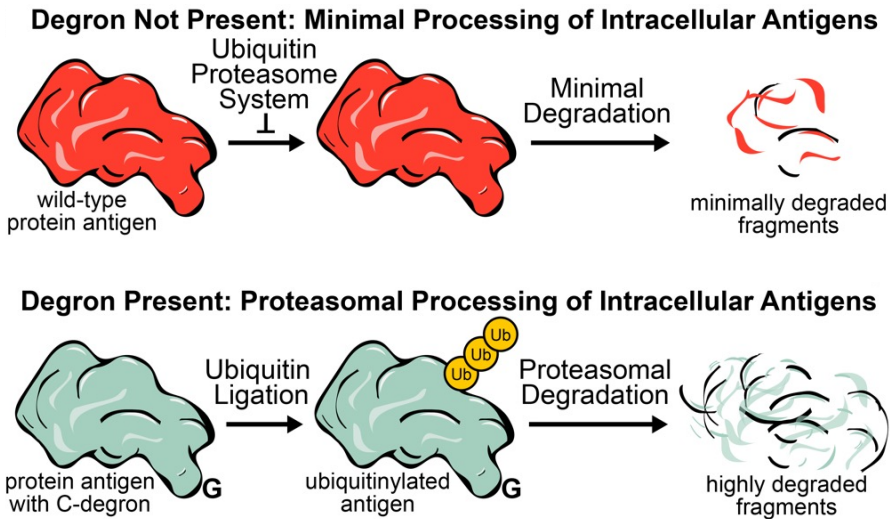


Figure 3-17: **Effect of degron on antigen processing.** Antigens without a degron sequence undergo minimal antigen processing, while anitgens with a degron harness the ubiquitin proteoasome degradation system to facilitate processing and presentation.

1986, it was noted that a lone N-terminal residue can affect the proteasomal stability of a peptide, now known as the 'N-end rule' [177]. Recent studies have shown how both N- and C-terminal motifs can affect proteasomal degradation [178–183]. In the context of antigen processing and presentation, the C-terminal is more critical than the N-terminal due to the trimming mechanisms [184, 185]. Unfortunately, the lack of tools to predict degrons has not been able to drive vaccine design using this approach.

In this work, we predicted and optimized the degron sequences, especially C-degrons, to tune antigen processing and presentation. We validated sequences with a wide range of proteasomal stability in *in vitro* experiments.

3.4.1 Data set

We used protein stability data obtained by perturbing the C-terminal and N-terminal motifs to predict degron activity [183, 186]. These data sets were generated by using libraries of DNA-encoded sequences of human proteins, conjugated to the C-terminus of green fluorescent protein (GFP). Flow cytometry analysis of the degradation activity of these DNA constructs, observed from the fluorescence intensity of the GFP-

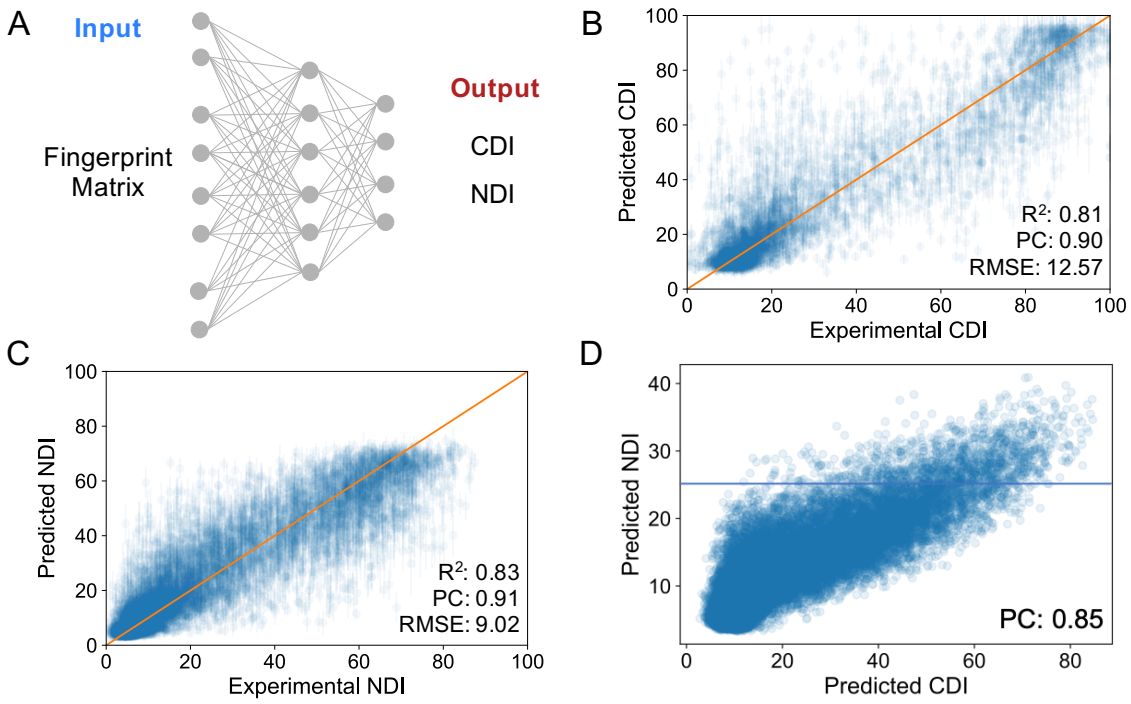


Figure 3-18: **Machine learning of degrons.** A. ML Framework with fingerprint matrix input, and CDI or NDI as output. Parity plots with performance metrics on test set for B. CDI and C. NDI. D. Joint plot of CDI and NDI for 100,000 randomly sampled sequences.

containing cell populations, was performed. The intensity plot was distributed into 4 numeric 'bins' (e.g., bin_1, bin_2, bin_3, and bin_4) with nearly equal cell population. The fraction of cell populations in the lower bins (e.g., bin_1) indicated more degradation activity; while the fraction of cell populations in the higher bins indicated less degradation activity (e.g., bin_4). Each cell was sequenced for the peptides, and for every peptide, the count was maintained across all the bins. Ultimately, based on the counts, the probability of the presence of the respective peptide was calculated for each bin.

3.4.2 Machine learning

To simplify the bin probabilities, we developed an aggregate scoring method to represent the proteasomal stability of a single peptide, instead of using individual bin populations. We refer to this aggregate score as the C-terminus degron index (CDI)

for the data set with modifications at the C-terminus, or N-terminus degron index (NDI) for modifications at the N-terminus. This score represents the extent of proteasomal degradation activity, in which a low CDI/NDI indicates more degradation, but a high CDI/NDI indicates less degradation. Quantitatively, we calculated this value using a linear combination of bin probabilities and exponentially increasing coefficients (e.g., 0, 1, 10, and 100) -

$$CDI = bin_1 \times 0 + bin_2 \times 1 + bin_3 \times 10 + bin_4 \times 100 \tag{3.2}$$

We used CDI and NDI scores to develop a CNN model ensemble over peptide sequences represented as matrices of residue fingerprints (Figure 3-18). The model ensemble was trained, validated, and tested by splitting the flow cytometry data into three subsets: 60:20:20. On the respective held-out test data sets, we obtained root-mean-squared error (RMSE) values in absolute units, CDI_{RMSE} 12.57, and NDI_{RMSE} 9.02. We analyzed overlapping N- and C-degron activity for shorter peptides (<25 amino acids), which revealed a Pearson's correlation of 0.85. This analysis indicates some crossover activity between N- and C-terminal degrons for shorter peptides, yet still permits validation through selecting sequences with varying CDI values but a constant NDI.

We used random sampling of sequences to query the prediction models, in order to understand the influence of individual amino acids and positions at the C-terminus on proteasomal processing. By randomly varying 10-residues from the C-terminus, while keeping the epitope and the N-terminus constant, we generated a total of 30,000 sequences. It was observed that these sequences had CDI score from 25 to 60, and there were unique trends associated with both higher and lower CDI scores. Specifically, we noted that glycine residues at -1 and -2 positions strongly contributed to low CDI scores, consistent with observations in the literature. However, it was noted that glycine at other positions did not have as much influence. Overall, we saw that alanine and cysteine-containing peptides had lower CDI scores, while the presence of arginine and valine at certain positions resulted in a lower CDI score. For higher CDI

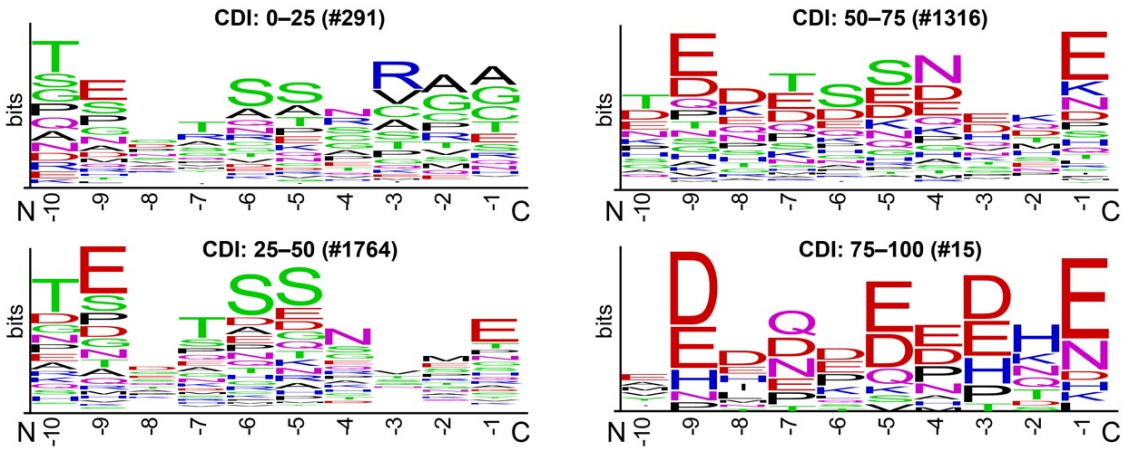


Figure 3-19: **Sequence logos of degrons.** Sequence logos of recurrent amino acids at corresponding positions, which were plotted across four quartiles of CDI values (0–25; 25–50; 50–75; and 75–100). The number of sequences represented are shown in parentheses.

scores, aspartic acid, glutamic acid, and lysine-containing peptides, and tryptophan, phenylalanine, isoleucin and leucin at specific positions were the key contributors.

In a parallel approach, we analyzed the amino acid frequency distribution using sequence logos for the random sequences with CDI in different quartiles, 0–25, 25–50, 50–75, and 75–100 (Figure 3-19). A strong influence of the C-terminal residues with the CDI scores was noted with the presence of specific residues in different quartiles. One such example is the presence of alanine, glycine and argninine in the first quartile as residues contributing to low CDI scores.

3.4.3 *In vitro* validation

Intracellular delivery is essential for evaluating proteasomal degradation of degron sequences. We achieved intracellular delivery using two anthrax proteins (Figure 3-20): protective antigen (PA) and the N-terminus of lethal factor (LFN) [187]. We used these proteins to evaluate seven peptides derived from chicken ovalbumin (OVA **1**–**7**), which were synthesized and conjugated to the C-terminus of LFN using sortase-mediated ligation [188]. Through enabling robust intracellular delivery, these constructs permit the characterization of relative degron activity based on CDI values

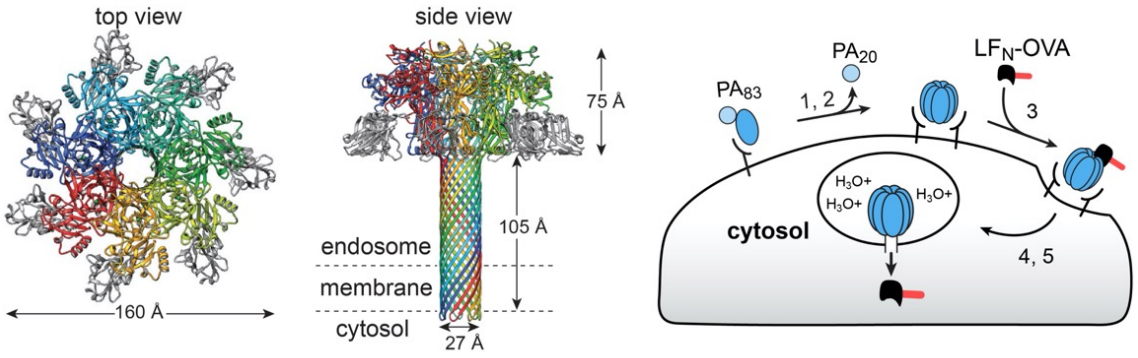


Figure 3-20: **Anthrax delivery system.** Atomic model of the active protective antigen (PA) pore, and cartoon illustration of the translocation mechanism from the PA and N-terminus of lethal factor (LF_N).

for OVA **1**–**7**.

We evaluated degron activity using T cell proliferation assays (Figure 3-21). We envisioned the extent of proliferation would reflect proteasomal degradation and, in turn, antigen processing and presentation. For these assays, mouse splenocytes (C57BL/6) were treated with the PA/LFN-OVA **1**–**7** constructs, followed by co-incubation with CFSE-stained T cells (OT-1). After 24 h, activation of the T cells was measured by flow cytometry based on upregulation of CD69 and CD137 markers. After 72 h, T cell proliferation was assessed by flow cytometry based on dilution of the CFSE dye. This study revealed a correlation between the CDI value and T cell proliferation activity, which suggests broader opportunities to tune immune recognition of molecular vaccines through increasing or decreasing degradation activity.

3.5 Prediction and optimization of flow synthesis

Organic reactions are arduous processes, usually involving multiple steps and necessitating labor-intensive and numerous rounds for optimization [189]. Automation of these have been explored using flow chemistry which offers better reaction outcomes, along with higher productivity and reproducibility, relative to batch methods [190, 191]. Some of the recent works have shown the development of a modular synthesis platform for synthesis and purification of small molecules using building blocks

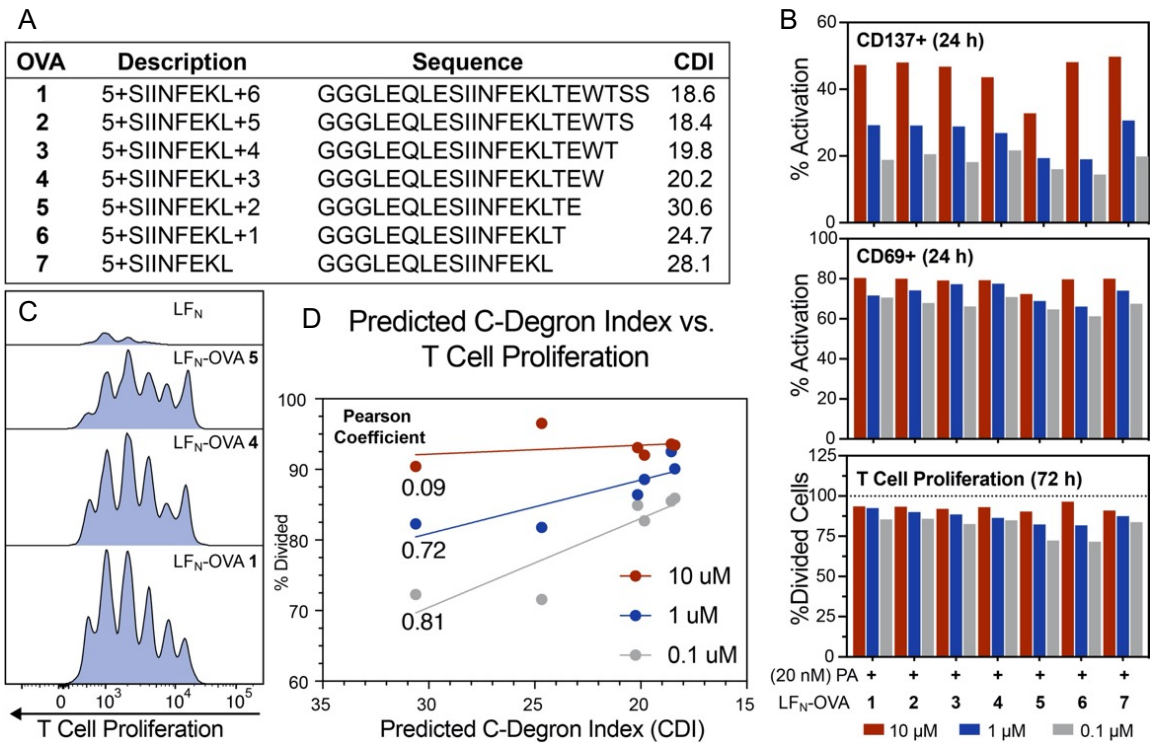


Figure 3-21: ***In vitro* validation.** A. List of peptide sequences conjugated to LFN:OVA 1–7. B. Flow cytometry analysis of T cell activation after treatments with PA (20 nM) and LFN-OVA 10, 1, and 0.1 uM): CD137+ and CD69+ (24 h); T cell proliferation (72 h). C. Representative T cell proliferation graphs from treatment with PA (20 nM) and LFN-OVA (1 uM). D. Correlation between predicted C-degron index (CDI) vs. T cell proliferation.

for Suzuki-Miyaura cross-couplings [192, 193]. In another instance of flow chemistry, a compact, multi-purpose benchtop synthesis system was developed to address several chemical reactions [189].

Advances in data science and artificial intelligence have transformed the heuristic-driven, trial-and-error intensive organic synthesis process 13-16. Combining state-of-the-art algorithms with robotics, automated systems that plan, execute and evaluate new experiments have been developed [194]. Specialized chemistry software, such as ChemOS [195], Chemputer [196], Ada [197] and Phoenics [198], have been developed to aid easy interfacing with hardware for chemists, along with automating the process. Other works around retrosynthesis [199], and optimization of reaction conditions have approached synthesis prediction from a different perspective [200, 201]. Despite all

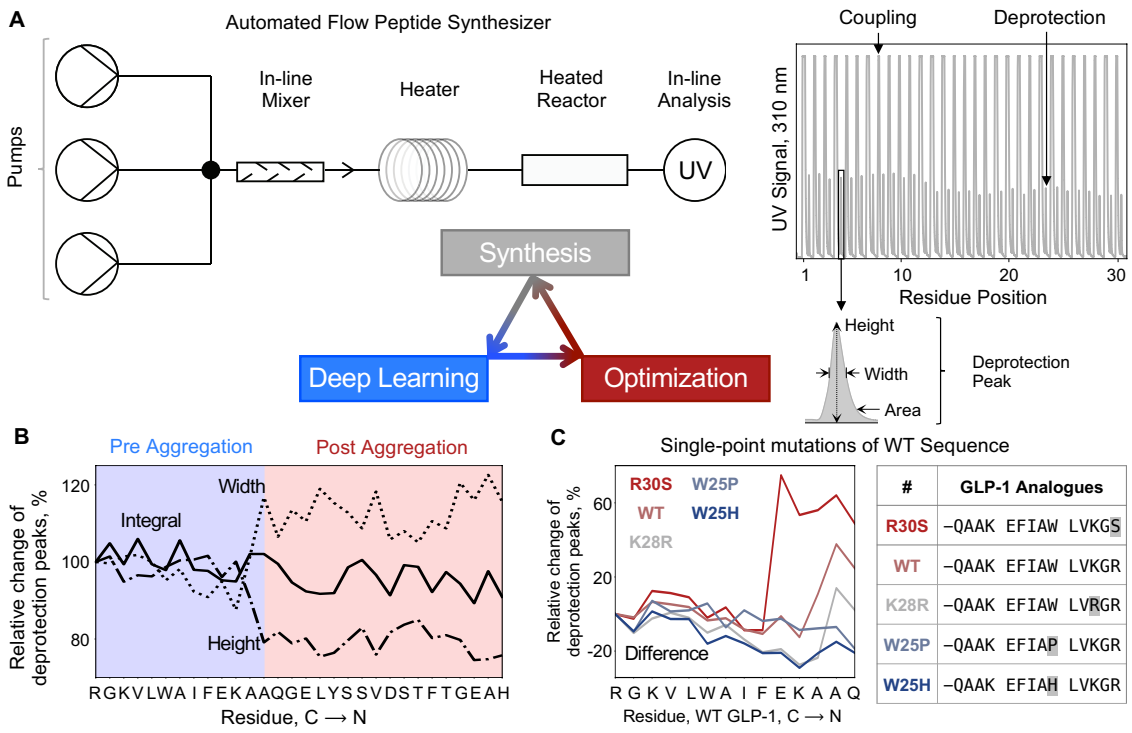


Figure 3-22: **Overview of flow synthesis prediction, optimization, and validation loop.** A. Schematic of data acquisition from the experimental setup. B. Per-residue reaction yield of peptide sequence to train on. C. Single residue mutation-based sequence optimization for improved synthetic yield.

these efforts, the prediction and optimization of synthetic yield of a complete reaction, especially for an unknown reaction, remains as a challenge.

A major bottleneck in the development of computational methods lies in the access to high quality data sets[202]. The data acquired from reaction platforms is often proprietary, customized to the specific experimental set-ups, and sometimes irreproducible, and hence not easily available and not able to be used directly [203].

In this work, we predicted flow synthesis reaction yields for peptides, using high quality peptide synthesis data generated in-house (Figure 3-22) [204]. We used the predictive models to optimize sequences to prevent aggregation and other sequence-dependent events [205–210]. In addition to optimization, we used attribution methods to interpret the decision-making process in the sequence optimization.

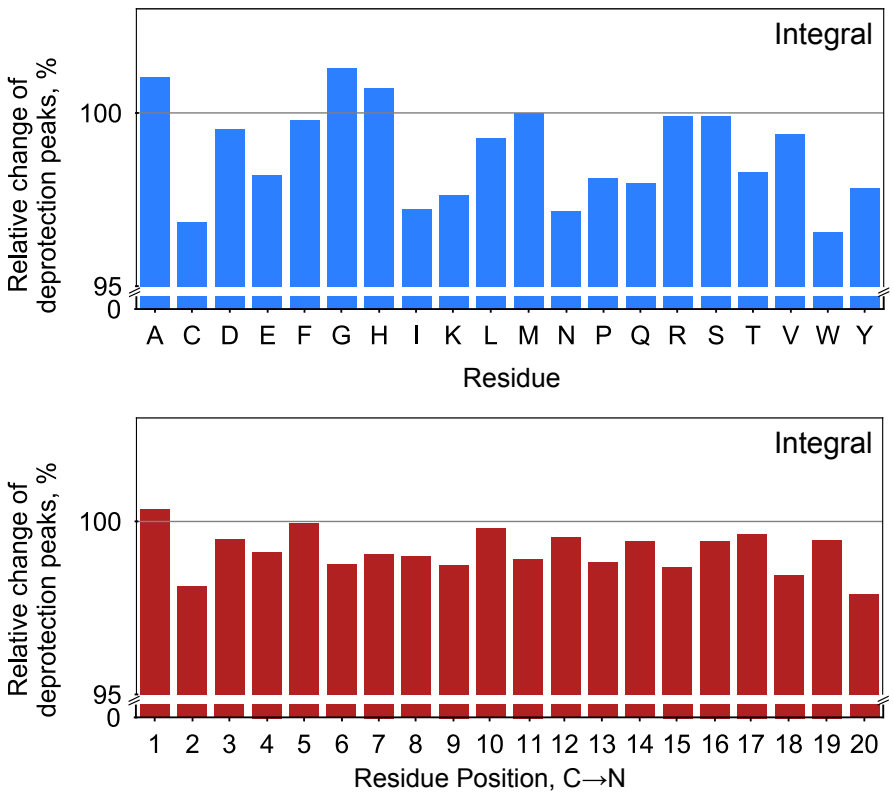


Figure 3-23: **Average synthetic yields per residue and position of sequence.** Integrals of deprotection peaks were analyzed and averaged for the addition of every incoming residue and the positions.

3.5.1 Data set

Synthesis data was generated using an automated fast-flow peptide synthesizer [121, 211]. Using an in-line UV-Vis detector, step-wise coupling and deprotection steps in the solid phase peptide synthesis process are monitored [212, 213]. The time-dependent data acts as a proxy for the step-wise or residue-specific synthetic yields, and also provides information for the complete reaction [214, 215]. Since the signal in the coupling step is saturated for every residue, we used the information for the deprotection step to quantify the yield.

The data set consisted of 35,427 individual synthesis steps, out of which 17,459 were unique [204]. Pre-synthesized sequence on the resin or pre-chain, incoming amino acid, and synthesis parameters, define each step. The area under the curve or integral, height, and full width at half the height for each deprotection step were

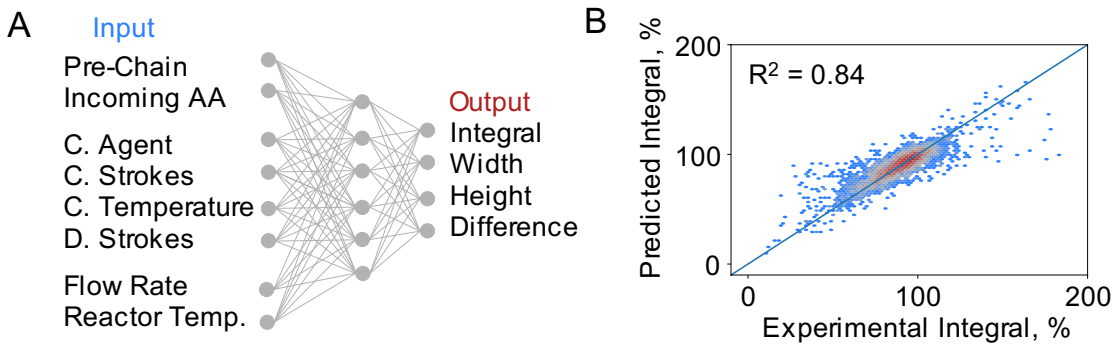


Figure 3-24: **Model framework and performance.** A. Schematic of the ML model with sequence data and synthesis parameters as input, and deprotection peak information as output. B. Parity plot of predicted versus experimental integral values for synthesis steps in the test data set.

used to quantify the yield. Apart from these values, the difference between width and height was used as an explicit representation for aggregation. Analyzing the average integral values per residue and for each coupling reaction along the peptide sequence, we note that there are huge variations in the efficiency at how each residue addition occurred (Figure 3-23).

3.5.2 Machine learning

Sequence data and synthesis parameters were used as inputs to the ML model (Figure 3-24A). For the sequence data, the pre-synthesized sequence on the resin was represented as a fingerprint matrix, while the incoming amino acid was featurized as a circular fingerprint [64, 70]. Each amino acid in the synthesis has side-chain protecting groups and fluorenylmethyloxycarbonyl (Fmoc) protection at the amide end. To represent the chemistry accurately, we used fingerprints of Fmoc and side chain-protected molecules for the incoming amino acids, and only side-chain protected molecules for residues in the pre-synthesized sequence. Key synthesis parameters were presented to the ML model as input values. Coupling agent and number of strokes were represented as categorical features, while coupling temperature, reactor temperature, and flow rate were represented as numerical values.

We trained multi-modal convolutional neural networks (CNN) over deprotection

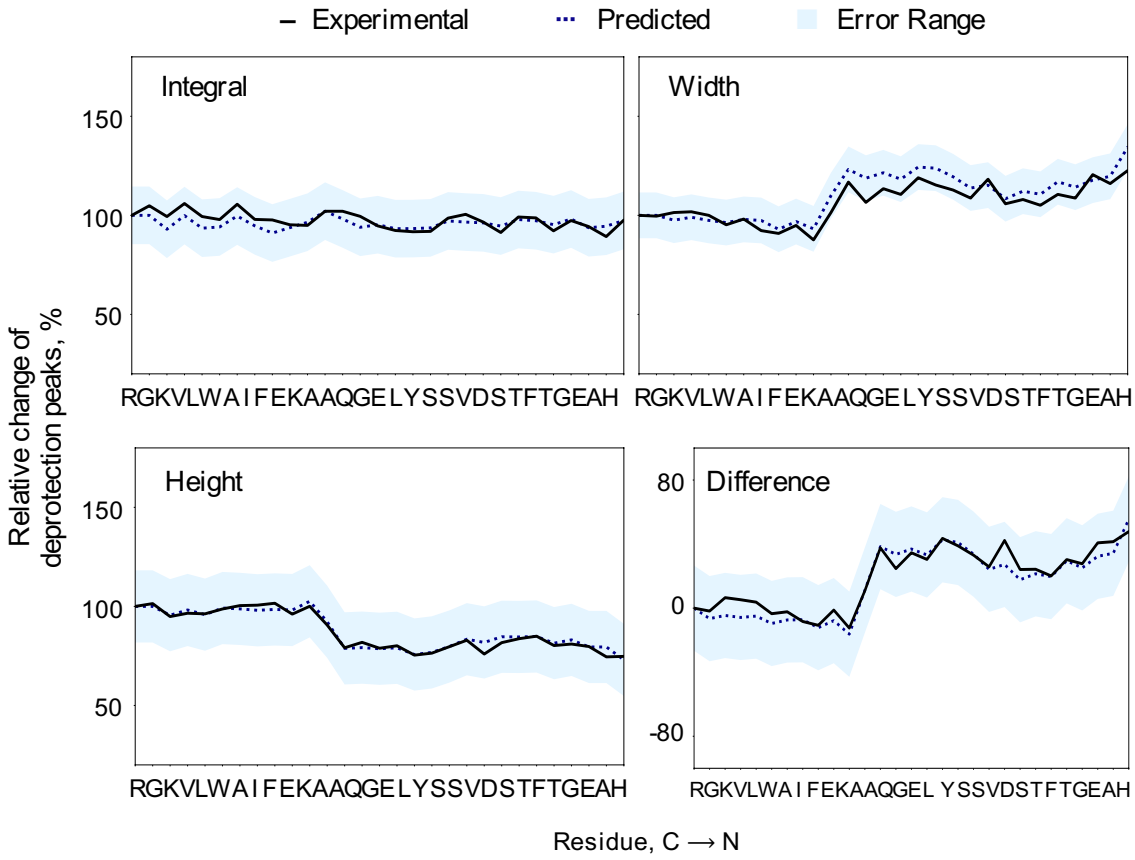


Figure 3-25: **Predicted and experimental traces for GLP-1 synthesis.** Experimental traces for integral, width, height and difference, have been overlaid for predictions with the uncertainty.

peak information as output (Figure 3-24B). The model architecture learnt individually over pre-chain, incoming amino acid, and synthesis parameter features, aggregated all the information, processed through a couple of fully-connected layers, and then made the final prediction of the output values. A 70:30 train: test random split of the data set was used for the model training. Evaluating on the held out test data set, for integral, width, and height, all predictions are within 0.1 root-mean-squared error. For GLP-1, a sequence not in the training data set, we noted that the predictions for the UV-Vis traces representing the synthetic yield nearly matched the experimental traces (Figure 3-25).

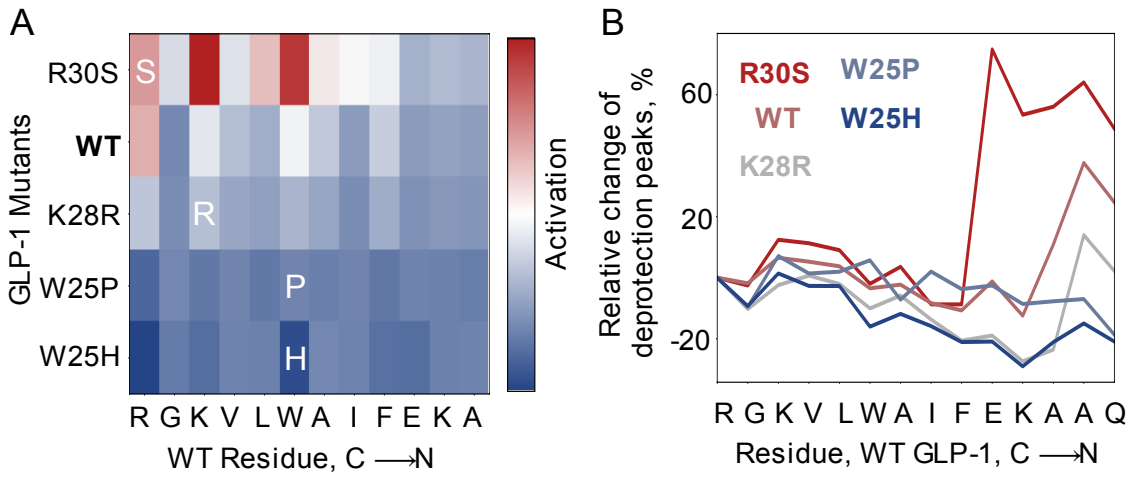


Figure 3-26: **Prediction and optimization of aggregation.** A. Heat map showing contribution of residues towards aggregation in wild-type and mutants of GLP-1. The more activated (red) a residue is, the stronger is its contribution towards aggregation. B. Experimental traces of difference for WT GLP-1 and optimized mutants.

3.5.3 Aggregation prediction and sequence optimization

We used the same model to predict the difference of width and height, representative of sequence-dependent events, referred to as aggregation in the synthesis (Figure 3-24). The model was able to predict aggregation in the held-out test data set with an 0.13 RMSE. For GLP-1, the model was able to correctly identify the aggregating event at Ala18 (A18) (Figure 3-25).

To interpret the decision-making process of the model using gradient class activation maps-based attribution introduced in [64], we trained a minimal model with sequence information input, and difference as the sole output. This approach enabled us to identify residues in the sequence and individual substructures that led to problems in the synthesis process. One of the key outcomes from this analysis was the realization that the residue most responsible for aggregation could be far away from the aggregating step, and sometimes in the initial steps of the synthesis, contrary to the contemporary intuition.

Single-point mutation was used to optimize sequences to reduce aggregation in the flow synthesis process (Figure 3-26). By brute-forcing through all possible mutations, we predicted and ranked how each mutation would affect the aggregation event. From

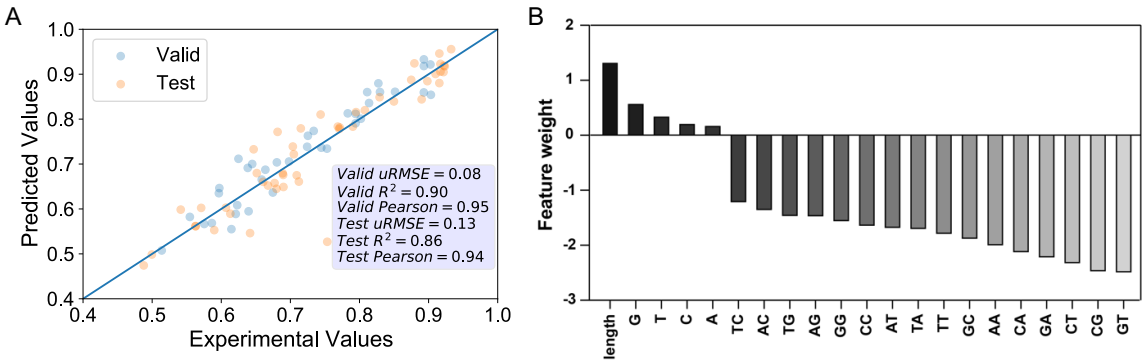


Figure 3-27: **Prediction of PNA synthesis.** A. Parity plot of predicted and experimental integral values for validation and test data sets using ridge regression. The inset text box notes the model performance metrics. B. Linear coefficients of ridge regression model for different features.

the list of mutants, we selected four sequences, with varying yield, to synthesize and validate the model predictions. We noted that the experimental traces of all the sequences were within error range of the predicted traces. The reduction of aggregation also resulted in higher purity of the final product, and helped in ease of downstream processing of the sequence.

We extended our aggregation analysis to include the proteins in the Protein Data Bank (accessed on April 17, 2020) [216]. In total, we analyzed 8,441 natural proteins with less than or equal to 50 amino acids. We found that 45% of the sequences were found to be aggregating, or having at least one synthesis step where the difference is greater than 20. Similar to our observation for GLP-1, we noted that residues closer to the C-terminus, where the synthesis starts, are most likely to cause aggregation than others in subsequent positions. Also, residues with aromatic and bulky side-chain protecting groups are more responsible for aggregation, as compared to others. Our analysis is in line with reports in the literature where hydrophobic amino acids have been flagged to cause aggregation [205, 208].

3.5.4 Extension to peptide nucleic acid flow synthesis

Peptide nucleic acid (PNA) platform combining the anti-sense oligonucleotide therapeutics [217–221], with delivery peptides, is gaining a lot of traction to address gene-

specific therapies [222, 223]. This platform offers enhanced chemical and thermal stability, apart from high target-specificity [224]. Recently, we developed a fully-automated synthesizer, on the lines of the automated flow peptide synthesizer [121], to enable rapid, reproducible PNA synthesis [225, 226].

A PNA synthesis data set using in-line UV monitoring was generated, similar to [204]. To normalize the deprotection traces, a 3-mer Lys motif was inserted at the C-terminus of the PNA sequences, and integrals of all peaks were normalized to the average of the integrals of the lysine motif. The final data set consisted of 239 unique pre-chain and incoming nucleotide combinations.

Given the significantly less amount of data for PNA synthesis, as compared to peptide synthesis, we adapted both the feature representations and ML methods. The pre-chain and incoming nucleotides were represented using one-hot encodings denoting the presence or absence of the 1-mer and 2-mers. A total of 21 different input features were used - four 1-mers for the 4 nucleotides, sixteen 2-mers for the sixteen possible combinations, and the length of the pre-chain. The normalized integral value was used as output to represent the synthetic yield of the step.

A variety of ML model architectures, ranging from linear, random forest, gradient-boosted tree models to Gaussian processes, were used to train over the PNA synthesis data set. For the training and evaluation, we split the data set as 60:20:20 for training, validation and test, respectively. Ridge regression was noted to be the best model with the lowest RMSE, 0.07, and highest Pearson's correlation, 0.97, on the held out test data set (Figure 3-27).

We analyzed the coefficients of the ridge regression model to understand feature importance. In a linear model, such as ridge regression, the coefficients directly correspond to the positive and negative importance of each input feature. Here, we noted that length and the presence of different PNA monomers are key to the success of the synthesis step, more than any of the dimer features.

To validate the model performance, PNA sequences were generated and experimentally synthesized. For a set of six random sequences, consisting of three 10–mers, one 6-mer, one 14-mer and one 18-mer, we observed that the predicted traces were

similar to the experimental traces, thereby indicating the prediction accuracy of our ML model. Additionally, we also predicted synthetic traces for all potential antisense PNA sequences targeting DMD [227]. We selected three PNA sequences with varying predicted yields, experimentally synthesized them, and compared with the predictions. In line with the predictions, we noted that the easy-to-synthesize sequence could be obtained with high purity, while we faced significant challenges and could obtain only trace amounts for the difficult-to-synthesize sequence.

3.6 Similarity computation, machine learning and optimization of glycans

Glycans are one of the most diverse class of macromolecules having non-linear topologies and a wide range of monomer and bond chemistries [228]. In glycan databases, such as GlycoBase [229] and GlycomeDB [230], we noted more than a thousand unique carbohydrate monomers. This number is between one and two orders of magnitude greater than what is observed in other biological macromolecules - polynucleotides, with 4 monomers, and proteins, with 20 naturally occurring monomers. The monomers can be connected with other monomers at one or more of the carbon atoms in the ring or side chains, and by a variety of bonds with different stereochemistries, further contributing to the diversity of the chemical space [231, 232].

A number of structural, metabolic, immune and regulatory functions are driven by glycans [233–236]. Lectins, or glycan-binding proteins, play a significant role in the upregulation and downregulation of these functions [237]. Glycan profiles have been shown to correlate with numerous conditions, such as diabetic retinopathy [238], cancer [239], Alzheimer's disease [240, 241], amongst several others. However, due to limited technological advancements, widespread study of the effect of glycans on the human body, or glycomics, is lacking compared to its biological macromolecule counterparts - genomics and proteomics [242–245].

Given the sheer complexity of these macromolecules, relatively limited experi-

mental attempts have been made to study them in depth [246, 247]. Amongst these attempts, the use of glycan arrays for profiling binding affinity of diverse lectins to the printed glycans is most prominent [248]. This approach has led to large data acquisition attempts, and ultimately led to the founding of the Consortium of Functional Glycomics (CFG), which hosts a number of experimental data repositories [249]. In addition to the study of lectin-glycan binding, consolidation efforts have resulted in databases of functional glycans with immunogenicity and taxonomic labels [250].

Similar to the experimental attempts, the complexity of the glyan chemical space and the lack of extensive data sets have resulted in limited computational attempts. Similarity computation of glycans has been attempted using tree-based methods [51, 53, 251, 252], and more recently using a BLOSUM-inspired GLYSUM evolutionary matrix-based motif alignment approach [34, 42]. While the former class of methods are more mathematical and do not have a solid chemical basis for similarity computation, the latter is limited to linear glycans alone.

Machine learning studies directed towards glycans, and lectin-glycan interactions have used non-chemical representations. A large number of works have used text-like representation, such as 'glycowords' in [34] and n-grams for monomer, dimer and trimer structures in [253]. Sub-tree mining of the glycan graph has also been used for prediction tasks [254, 255]. Despite using powerful model architectures, the lack of a chemistry-informed representation has limited the performance and interpretability of the models.

In this work, we developed a chemistry-informed graph representation for non-linear macromolecules, such as glycans, using which we performed similarity computation, unsupervised and supervised machine learning for immunogenicity and taxonomy-prediction tasks (Figure 3-28). We used the same framework to predict lectin-glycan binding affinity, and employed graph-based genetic algorithm (GBGA) optimization to generate new glycans with high predicted binding affinity and low probability of human immunogenicity.

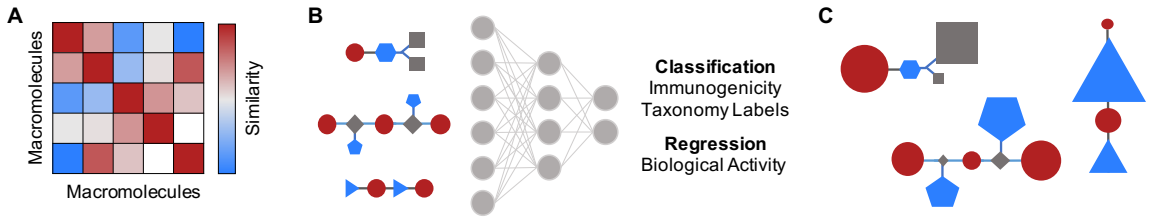


Figure 3-28: **Overview of computational framework for non-linear macromolecules.** A. Representative heatmap of pair-wise similarity computation of a library of macromolecules. B. Schematic of ML model with macromolecule graph input, and output for a variety of classification and regression tasks. C. Attribution highlighting the key monomer nodes corresponding to their importance for a particular task.

3.6.1 Data set

We downloaded a data set of 19,299 glycans from GlycoBase (accessed on November 2, 2020) [229]. The file listed the glycans in a nested-bracket string format, with species and immunogenicity labels. The bracket system closely followed the SMILES convention [35], where the branches were within distinct brackets, with sub-branches in the nested brackets. After data curation and removal of invalid data points and glycans for which no monomers could be identified, we had 19,147 glycans.

A custom parser was developed to parse the glycan strings to intermediate text files and final NetworkX graphs [256]. The intermediate text files were designed as a generalized text file format for sequence and topology-defined macromolecules. These files had SMILES, MONOMERS and BONDS as individual segments, inspired by the Protein Data Bank (PDB) file format [216]. SMILES listed monomer and bond codes, followed by the SMILES representation of the molecule. MONOMERS listed the position of each monomer in a 2D graph, and BONDS enumerated the connectivity between the monomers. These text files were then parsed into macromolecule graphs, with monomers as nodes and bonds as edges, both featurized using circular fingerprints.

3.6.2 Similarity computation and unsupervised learning

Pair-wise similarity of the glycan graphs was computed using GED for a limited number of glycans, and graph kernels for the entire library (Section 2.2.2). The output of a graph kernel is a kernel matrix with similarity vectors, or individual row/column entries, representing the similarity of an individual glycan with the rest of the library. These vectors are a powerful representation as they capture the variation in both chemistry and topology of the graph across the library, and can be readily used for downstream analysis, motivated by multi-dimensional scaling with linear and non-linear dimensionality reduction methods [257].

Unsupervised learning over the individual similarity vectors using dimensionality reduction and clustering provided insights into the sub-families of glycans (Figure 3-29). The immunogenic and non-immunogenic glycans clustered in nearly distinct spaces on the 2-component UMAP projection [79], with the former being on the fringes, and the latter forming the core. Similarly, when we visualized glycans by their taxonomic labels, bacteria were at the core, eukarya were spread out from the core, and viruses were at the fringes. We benchmarked our dimensionality reduction approach against t-SNE [78] and PCA [258], noting that t-SNE was better at capturing local structure and PCA the global structure, while UMAP performed better in separating the glycans at both levels.

3.6.3 Supervised learning

We performed classification of glycans for immunogenicity and different levels of taxonomy using five distinct graph neural network model architectures - Weave [259], Message Passing Neural Networks (MPNNs) [260], AttentiveFP [261], Graph convolutional networks (GCN) [262], and Graph Attention Networks (GAT) [263], implemented in Deep Graph Library [264]. We benchmarked fingerprint and one-hot encoding representations, and evaluated the model architectures thoroughly. We randomly split the data set in 60:20:20 training, validation and test sets. Hyperparameter optimization for the models was done using SigOpt [73]. The top five hyperparame-

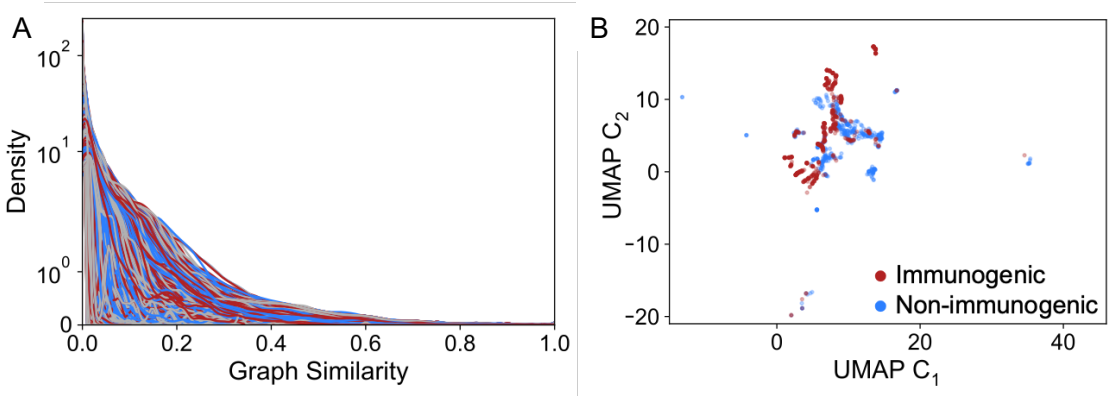


Figure 3-29: **Similarity computation and unsupervised learning of glycans.** A. Overlay of similarity vectors normalized to the maximum value of the vector for all the glycans in the data set. B. 2-component UMAP and coloring of glycans by immunogenicity labels.

ter sets were re-run with five different weight initialization seeds to obtain the final models.

For classification of immunogenic versus non-immunogenic glycans, we obtained state-of-the-art results, achieving a receiver operating characteristic-area under curve (ROC-AUC) score of 0.99 on the held-out test data set (Figure 3-30). Out of eight glycan classification tasks reported in the literature, we achieved state-of-the-art results in four, and comparable results in the remaining tasks [265]. In addition to the classification tasks, we also performed regression over anti-microbial activity of peptides, in order to evaluate the models over continuous-valued tasks for non-linear macromolecules.

3.6.4 Attribution

We used integrated gradients (IGs) [266] and Input x Grad (InpGrad) [267] for the attribution analysis of the GNNs - Weave, Attentive FP and MPNN. The model architecture selection was done to have one of each type of architecture – Weave (graph convolution), AttentiveFP (graph attention), and MPNN (message passing).

In the following formulation for attribution, the macromolecule is represented as a graph, $\mathcal{G}(V, E)$, where V represents monomers at vertices/nodes, and E represents connecting bonds at the edges.

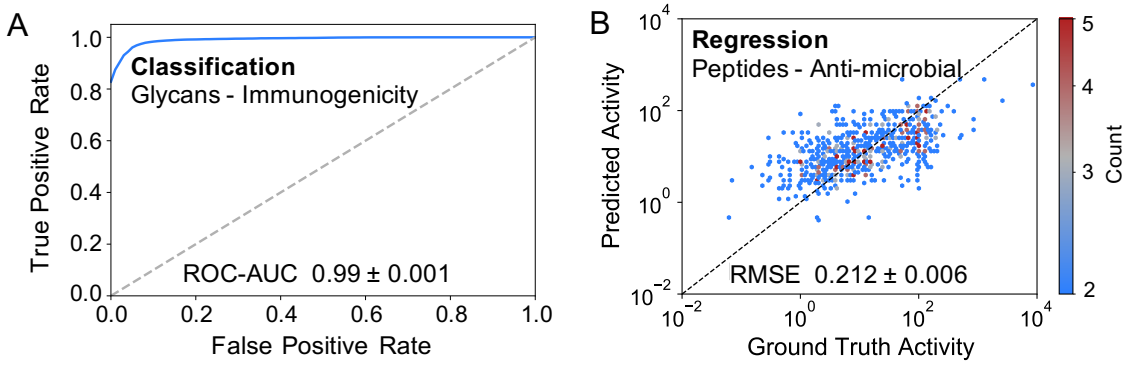


Figure 3-30: **Supervised learning.** A. ROC-AUC curve for the glycans in the held-out test data set for immunogenicity classification of glycans. B. Parity plot of predicted and ground truth anti-microbial activity for regression over peptides.

IGs interpolate between the input graph and a baseline graph with zero-valued features, accumulating the gradient values for each node (Equation 3.3). The notation follows [268].

$$\mathcal{G}_A = (\mathcal{G} - \mathcal{G}') \int_{\alpha=0}^{1} \frac{dy\,(\mathcal{G}' + \alpha\,(\mathcal{G} - \mathcal{G}'))}{d\mathcal{G}} d\alpha \tag{3.3}$$

InpGrad is the element-wise product of the input graph and the gradient.

$$\mathcal{G}_A = \left(\frac{d\hat{y}}{d\mathcal{G}}\right)^T \mathcal{G} \tag{3.4}$$

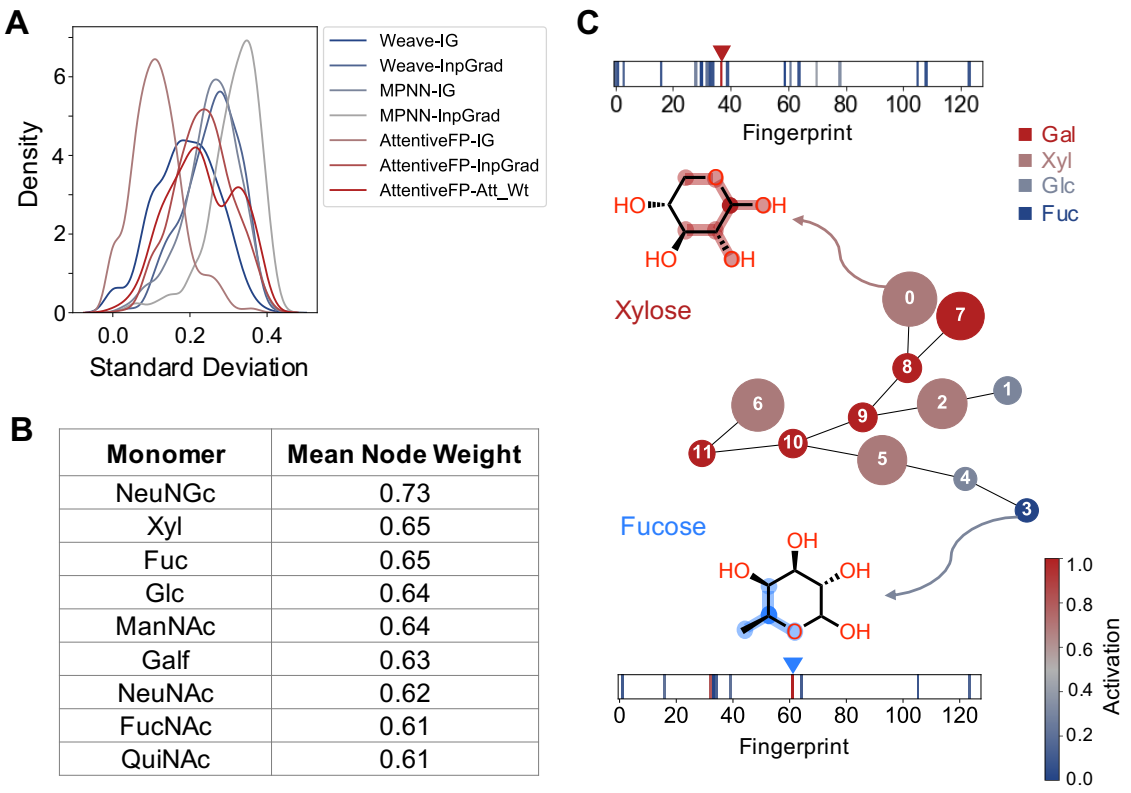
For the attention-based GNN, AttentiveFP, in addition to IGs and InpGrad, we evaluated attribution using attention weights, where the node attention weights are obtained by averaging over the attention scores of the adjacent nodes.

The node weights were obtained by multiplying the positive weights with the input fingerprint vectors, and then normalized to the maximum node weight to obtain the normalized weights-

$$\mathbf{n} = \sum_{nodes} \mathcal{G}_A^+ \mathcal{G}, \tag{3.5}$$

$$\mathbf{n_{norm}} = \frac{\mathbf{n}}{max(\mathbf{n})} \tag{3.6}$$

We used AttentiveFP-IGs for attribution analysis of glycans, identifying both key



Monomer	Mean Node Weight
NeuNGc	0.73
Xyl	0.65
Fuc	0.65
Glc	0.64
ManNAc	0.64
Galf	0.63
NeuNAc	0.62
FucNAc	0.61
QuiNAc	0.61

Figure 3-31: **Attribution analysis.** A. Distribution of standard deviation of node weights for different model architecture - attribution method combinations. B. List of key monomers contributing to immunogenicity of glycans, in descending order of importance. C. Visualization of weighted monomer nodes and substructures most responsible for immunogenicity in a representative glycan.

monomers and chemical motifs therein that contribute the most to immunogenicity (Figure 3-31). To find the most optimal model architecture - attribution method combination, we used the axiomatic approach described in [269]. In this context, the implementation invariance axiom requires that the node weights obtained across different implementations of the attribution-architecture combination be similar, with the ideal standard deviation being zero. Plotting the distribution of standard deviations for all the different combinations, we observed that for AttentiveFP-IGs, the mean of the distribution was the lowest. This observation can also be interpreted as different model instances consistently attributing the model prediction to the same nodes with similar importance.

Using the AttentiveFP-IGs combination, we analyzed the entire data set identi-

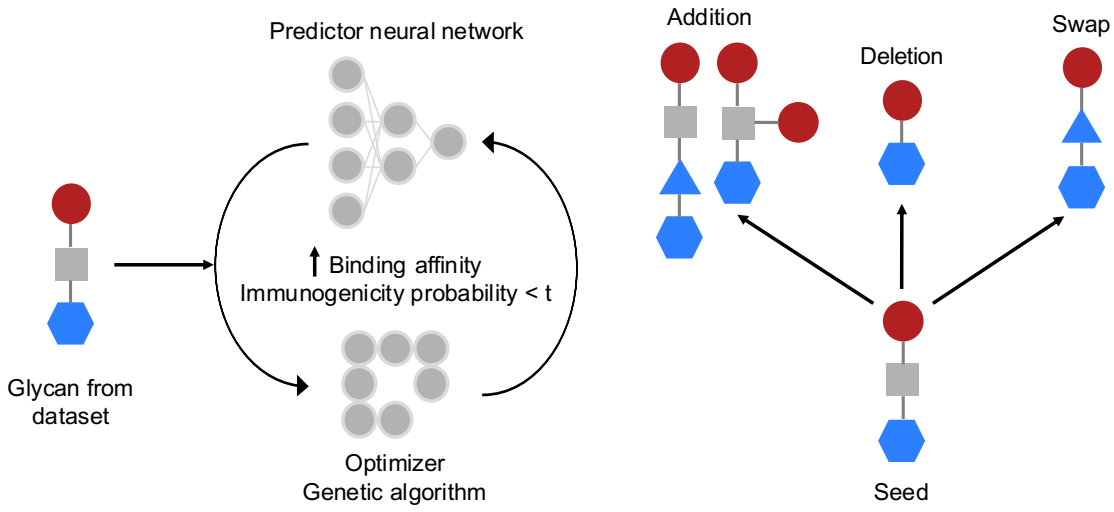


Figure 3-32: **Graph-based genetic algorithm optimization workflow.** A seed glycan is randomly mutated using a predictor-optimizer loop, increasing the predicted binding affinity, while keeping the immunogenicity probability below a threshold, t.

fying key monomers - N-glycolylneuraminic acid, xylose, and fucose - responsible for immunogenicity in glycans. For a single glycan, we visualized the key monomers that contribute towards immunogenicity. In addition to monomers, we also showed the substructures, such as the one centered on the anomeric carbon of xylose, that are key to the immunogenic activity of the glycan.

3.6.5 Optimization of lectin-glycan binding affinity

We applied the machine learning framework for prediction of binding affinity of glycans to lectins, and extended it further by adding a graph-based genetic algorithm (GBGA) optimization workflow. To demonstrate the application, we used mammalian glycan array data from the CFG for dendritic cell-specific ICAM-grabbing non-integrin (DC-SIGN) as the use case [248].

The predictor model was trained using glycan graph representations against corresponding binding affinities, represented as relative fluorescence units (RFU). All replicates of glycan-RFU pairs in the data set were used for the training, and the data set was split as 60:20:20 by the glycan for training, validation and test sets. Since RFUs had a right-skewed distribution, log transformed RFU values were used as out-

puts in the model training. An AttentiveFP GNN model ensemble was trained due to its prior success for classification and regression tasks of non-linear macromolecules in [81]. The ensemble had a root-mean-squared error of 0.782 of the standard deviation of the training data set, outperforming other results reported in the literature [253, 270].

The GBGA framework optimized a given seed glycan against an objective function for a specified number of iterations (Figure 3-32). In this case, the objective was increasing predicted binding affinity and keeping the immunogenicity probability under 0.5. In each iteration, the seed glycan undergoes a random mutation - insertion, deletion or swapping of a node, with the corresponding correction to its bonds and branches. The affinity of the mutated glycan is predicted, and if the affinity is greater than that of the seed glycan and the immunogenicity probability is less than the threshold, then the mutated glycan is used as the seed glycan for the next iteration. For DC-SIGN, the optimization was performed for the top fifty glycans, in descending order of binding affinity. We developed the framework as a generalized toolkit applicable to the single-value optimization of any macromolecule.

To select high affinity, chemically diverse glycans from the list of optimized glycans, we used similarity computation and dimensionality reduction techniques, similar to Section 3.6.2 (Figure 3-33). Using graph kernels, we obtained the pairwise similarity matrix. Dimensionality reduction of the similarity vectors using 2-component UMAP transformed the glycans into a more visually accessible projection space. We used k-means clustering, weighted by the binding affinity to identify five glycan clusters [271]. From each cluster, we selected the glycan with the highest binding affinity. The selected glycans are chemically diverse, while having high binding affinity to DC-SIGN and low immunogenicity probability.

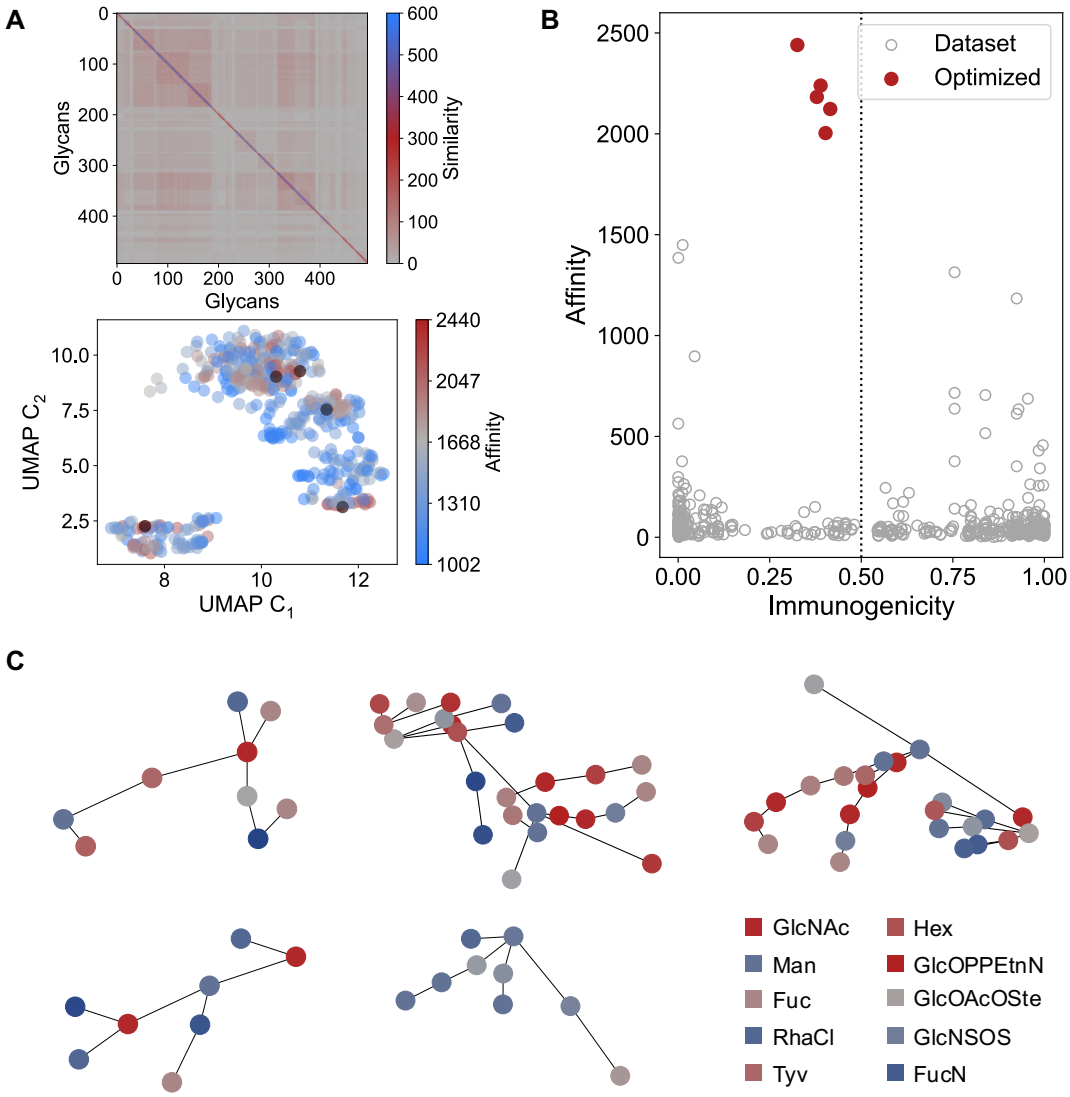


Figure 3-33: **GBGA for DC-SIGN-binding glycans.** A. Dimensionality reduction over the similarity matrix, and weighted k-means clustering for identification of high affinity and chemically diverse glycans. B. Plot of immunogenicity and affinity of glycans in the data set and the optimized glycans. C. Visualization of the optimized glycans with the legend of the common monomers.

Chapter 4

Applications to artificial macromolecules

4.1 Prediction, screening and validation of oligoethylene glycol-based Li-battery electrolytes

Quantitative attribution of the effect of molecular substituents have significant applications in a broad range of fields, ranging from catalysis to semiconductors [272–276]. Materials design necessitates these effects to be understood across different scales from the atomistic mechanism to the macroscopic property [277, 278]. For electrolytes, Edisonian approaches with thorough analysis of individual systems has improved our understanding, but *a priori* quantitative prediction remains a challenge [279–283]. To develop the next-generation of safe, easily processable electrolytes with high room temperature conductivity, comparable to liquid organic electrolytes or solid state ceramics, a model system and a quantitative model are both necessary [284–286].

For the model system, we chose the well-studied class of oligoethylene glycol-based electrolytes (Figure 4-1A) [287, 288]. Multiple studies have shown the enhancement of conductivity and viscosity by introducing substituents, copolymer blocks, small molecule additives [289–297].

Given the large number of possible interactions, it has been difficult to experimen-

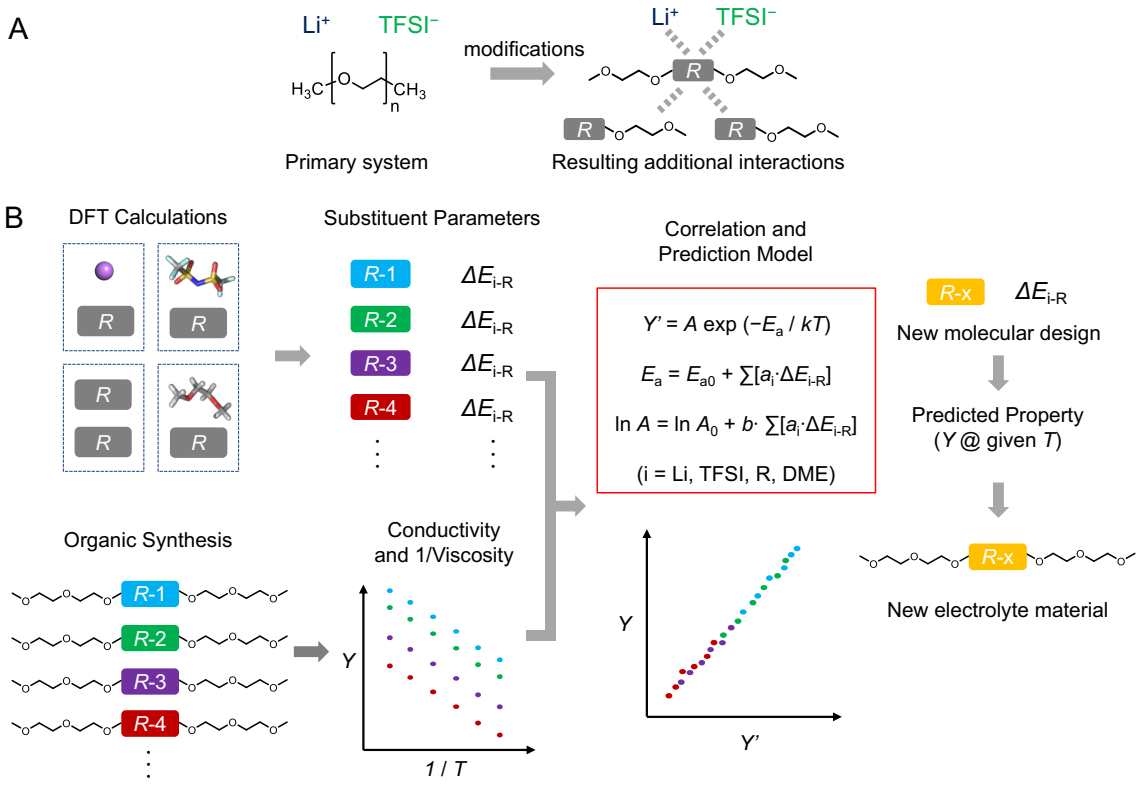


Figure 4-1: **Model system and overview of approach.** A. Interactions in OEG-electrolyte system with and without substituents. B. DFT simulations, experimental synthesis and characterization, modeling, and experimental validation framework.

tally screen several compounds or probe the mechanism of action [298]. Molecular dynamics simulations have been limited owing to the need for custom force fields and the amount of time required to simulate the necessary time scale [299–307]. Density functional theory (DFT) simulations have provided insights for smaller molecules, but have not been able to scale to necessary time or length scales [308–310]. The use of conventional ML approaches necessitate large data sets or accurate descriptors, both of which have been hard to find for the electrolyte systems of interest [36, 311–316].

In order to address these challenges, we used an enthalpy-entropy compensated Arrhenius equation with DFT computed pairwise interaction energies to model the effect of secondary substituents on the experimentally obtained temperature-dependent conductivity and viscosity (Figure 4-1B). We used this regression model to screen a small library of OEG-based electrolytes, out of which we synthesised and characterized three compounds with high, medium and low conductivity.

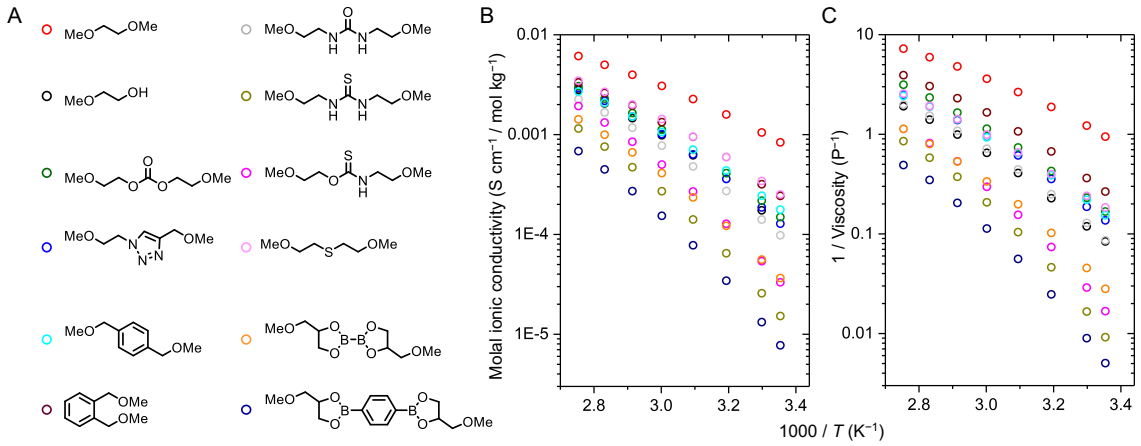


Figure 4-2: **Experimental data set of OEG-electrolytes with molecular substituents.** A. Different molecular substituents in the experimental data set. B. Experimental conductivity and viscosity at different temperatures.

4.1.1 Data set

We synthesized and characterized eleven OEG-oligomer electrolytes with different molecular substituents, and one without any substituent (Figure 4-2). The oligomer library constituted of substituents with a variety of chemical characteristics - hydrogen bond donors (triazole, urea, thiourea, thiourethane, hydroxyl groups) and acceptors (triazole, carbonate, urea, thiourea, thiourethane, hydroxyl groups), Lewis acids ((glycolato) diboron, benzenediboronic ester) and bases (carbonate, triazole, thioether, hydroxyl), and nonpolar aromatics (xylene). In addition to the following substructures, we developed a generalized procedure that can be used to design new substituents for OEG-based electrolytes.

The molal ionic conductivity and viscosity were measured across a range of temperature values. A constant lithium to oxygen ratio, 1/12, was maintained to minimize the differences between the substituents. Experimental observations show a remarkable difference in the conductivity and viscosity values, spanning approximately three orders of magnitude. The experimentally observed range of the molal conductivities of the OEG-based electrolytes, 0.0005–0.004 $Scm^{-1}/molkg^{-1}$ is consistent with the results reported in the literature, 0.0005–0.003 $Scm^{-1}/molkg^{-1}$ [317, 318]. We also confirmed that the conductivity and viscosity values for individual substituted OEG-

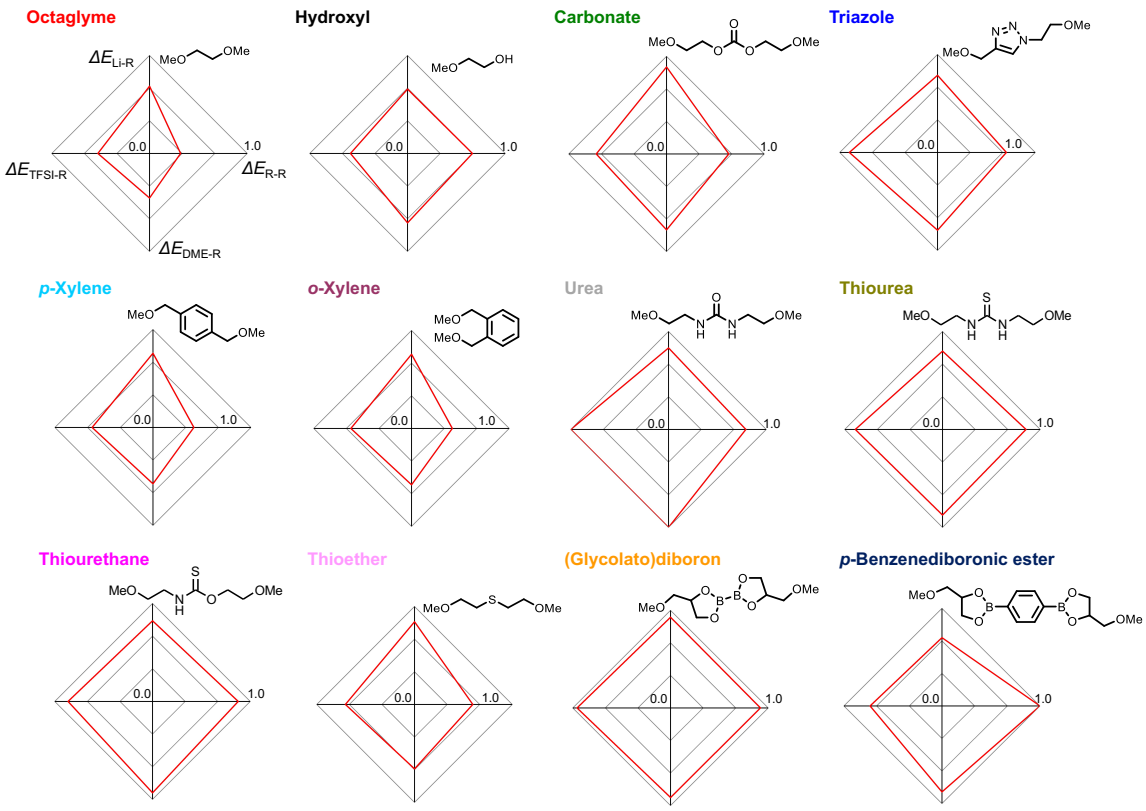


Figure 4-3: **DFT computed pairwise interaction energies for the molecular substituents.** Interactions with the oligoethylene chain, lithium cation, TFSI anion and self were considered.

electrolytes were inversely related, consistent with Walden analysis [317, 319, 320].

DFT calculations were performed to obtain pairwise interaction energies for four binary complexes - substituent with oligomer chain, lithium cation, bis (trifluoromethane) sulfonimide (TFSI) anion, and with itself (Figure 4-3) [321]. The substituent was represented using the smallest fragment separated by ether oxygen atoms from the OEG chain, and capped by methyl groups on both ends. These interaction energies were hypothesized to represent the effect of the substituents on the properties of the electrolyte, compared to the OEG-oligomer alone.

4.1.2 Arrhenius model formulation

The experimental conductivity and viscosity, and DFT computed interaction energies data sets were modeled using the Arrhenius equation: $\ln Y = \ln A - E_a/kT$, where

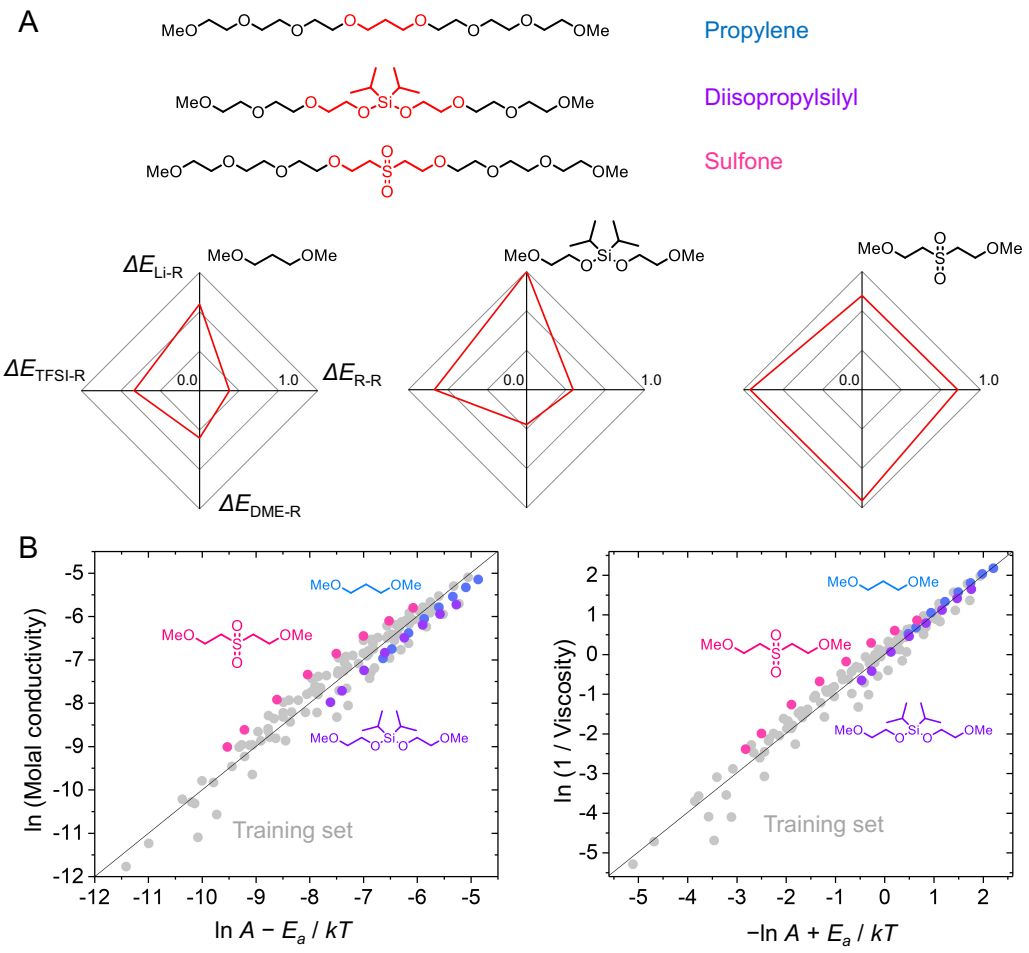


Figure 4-4: **Validation of new molecular substituents.** A. Substituted OEG-oligomer structures and pairwise interaction energies for the three screened substituents. B. Parity plot of model predicted and experimentally obtained conductivity and viscosity for the training data set and the new oligomers.

Y represents the experimental conductivity or viscosity, E_a is activation energy, A is pre-exponential factor, k is Boltzmann's constant, and T is temperature in Kelvin. (Figure 4-4). The formulation with the enthalpy-entropy compensation is -

$$\ln Y = \ln A_o + bC - \frac{E_a + C}{kT} \tag{4.1}$$

where C is $a_1\Delta E_{Li-R} + a_2\Delta E_{TFSI-R} + a_3\Delta E_{DME-R} + a_4\Delta E_{R-R}$, with a_i being the compensation coefficient for respective pairwise interaction ΔE_{j-R}, and j consists of the various components Li for lithium cation, $TFSI$ for TFSI anion, DME for a

dimer of the OEG-oligomer, and R for the substituent.

As the baseline case, we analyzed the E_a for the OEG-oligomer octaglyme alone, noting that the predicted value in for the conductivity model was 0.29 $\pm$ 0.02 eV close to the literature reported value of 0.34 eV for tetraglyme [317]. We also noted that the E_a for the OEG-oligomer alone was the lowest in the entire library, and the E_a considerably increased with the addition of the molecular substituents, with the highest being 0.65 eV and 0.66 eV for conductivity and viscosity of benzenediboronic ester. This increased E_a is also reflected in the increase in the value of A, which can be explained by the enthalpy-entropy compensation [322, 323].

In addition to the Arrhenius model, we also used the Vogel-Tammann-Fulcher (VTF) equation to model conductivity and viscosity accounting for the temperature-dependent deviation from the Arrhenius model: $\ln Y = \ln A - E_a/k(T - T_o)$, where T_o is usually considered to be $T - T_g$, with T_g being glass transition temperature [1, 324]. This modeling helped us obtain better fits for the conductivity and viscosity values. However, the E_a and A values were found to be within a small range across all the substituents, which can possibly be attributed to the addition of T_g which accounts for the variations in the properties.

4.1.3 Validation of new substituents

We manually generated a library of thirty-two different substituents, for which we calculated the pairwise interaction energies using DFT, obtained the conductivity and viscosity values using the Arrhenius model, and selected three substituents for validation (Figure 4-4). Amongst these three substituents, we predicted and successfully validated that the propylene-substituted OEG-electrolyte will have a higher conductivity, 0.86 $mScm^{-1}/molkg^{-1}$, than the OEG-electrolyte alone, 0.54 $mScm^{-1}/molkg^{-1}$ at room temperature, 298 K. For the other two substituents also, the predictions were nearly accurate as the experimentally observed conductivity and viscosity values, with the conductivity of diisopropylsilyl and sulfone substituted OEG-electrolytes being medium and low on the range of the conductivities in the original data set. We also predicted that fluorinated substituents for OEG-electrolytes would enhance the

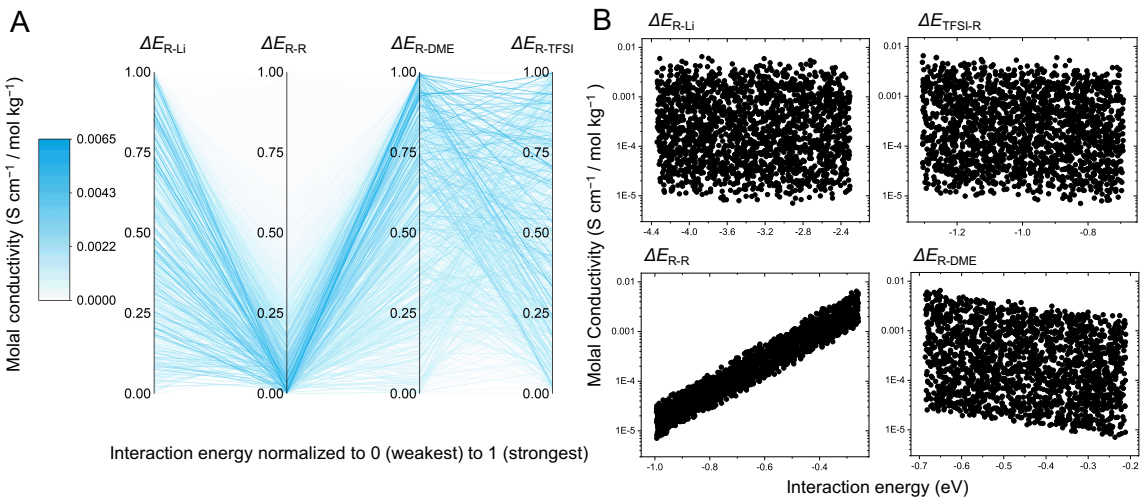


Figure 4-5: **Analysis of the effect of different interaction energies on conductivity.** A. Parallel coordinates obtained by random sampling of sets of different interaction energies and computing the conductivity. B. Correlation of individual interaction energies with conductivity.

properties, and this prediction was confirmed by another recent work for a similar system [325].

4.1.4 Parallel coordinate analysis

To understand how substituent interactions played a role in the conductivity, we used parallel coordinates to visualize interaction energies and conductivity. We generated sets of interaction energies for one thousand hypothetical substituents and computed the conductivity using the Arrhenius model. Each interaction energy was restricted to be within the range of the respective previously DFT-calculated energies. The predicted conductivity was colored on a scale of white to blue, depicting the range of low to high values. We also plotted the individual energies to the computed conductivity values.

High conductivity was found to be associated with low self-interaction energy, clustering at the bottom of the ΔE_{R-R} column in the parallel coordinates plot, and having a strong correlation ($R^2 = 0.86$) in the individual plot (Figure 4-5). The unexpected observation in this analysis was the negligible correlation of conductvity

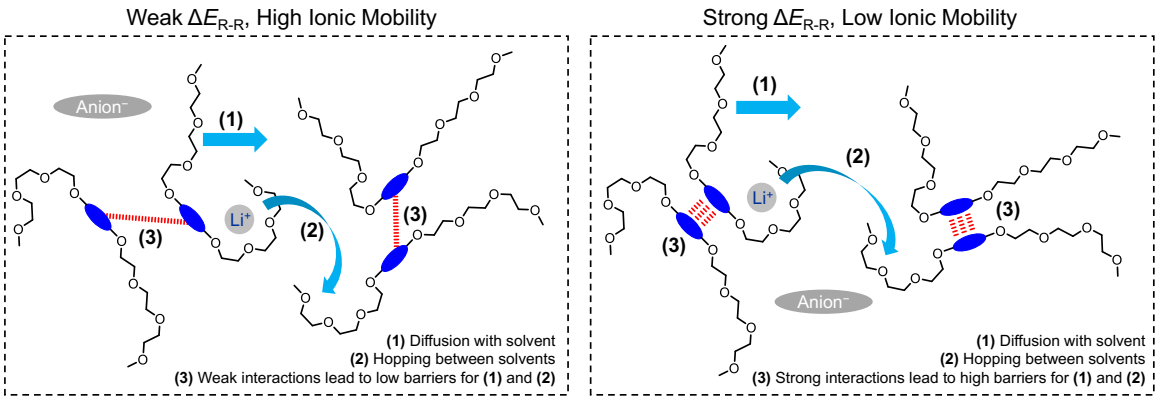


Figure 4-6: **Mechanism of action for the ionic conductivity.** Differences in the conductivity owing to weak and strong self-interaction energies.

with ΔE_{R-Li} and ΔE_{R-TFSI}, and the weak correlation with ΔE_{R-DME}.

Given these observations, we hypothesized that there is high ionic mobility due to the low self-interaction energy which manifest as weak interactions in the electrolyte (Figure 4-6). These weak interactions resulted in low energy barriers, aiding the diffusion of the oligomers and ease of hopping of lithium cations between the oligomers. On the contrary, strong self-interaction energies result in low ionic mobility.

4.2 Prediction, screening and validation of sustainable thermosets

The dramatic increase in the production and use of plastic, and the concurrent rise in waste production necessitates development of sustainable plastics [326, 327]. It is projected that the amount of plastic produced in the next 15 years will be more than the combined production of plastic all across human history [328]. Compounded by the long-term impacts of plastic waste, it is high time that we develop more sustainable and reusable plastics [329, 330].

Thermosets, at 20% constitute one of the largest categories of plastics manufactured [331–335]. These plastics are covalently crosslinked polymers that harden irreversibly upon curing. They are often used in extreme environments, such as in high-temperature, mechanically and chemically harsh settings. Given the irreversible

hardening and covalent crosslinking, it is difficult to recycle thermosets, thus requiring the disposal of the waste in landfills where they persist for long periods of time.

Cleavable comonomer additives (CCAs) have been introduced into the thermosets during crosslinking, within existing manufacturing protocols, to improve their lifecycle circularity [331, 332]. CCAs enable chemical deconstruction of thermosets, usually by a mechanism different from the ones possible in the chemical environment they are used in. The deconstructed products could be recycled, and used for making new thermosets. In a recent work, we showed how 7-10% cyclic olefin CCAs with silyl ether groups could copolymerize with norbornene derivatives to form pDCPD (polydicyclopentadiene), a widely used thermoset [336]. Through heuristic-driven design of CCAs, we were able to increase the glass transition temperature, T_g, necessary for high-temperature applications, while ensuring deconstructability of the thermosets. While this work is a proof-of-concept for the use of CCAs in the development of sustainable thermosets, the identification of new CCAs remains a key challenge.

Data-driven approaches have helped in accelerating materials discovery and synthesis for a wide range of applications. Polymer data sets from PoLyInfo [337] and PI1M [338] have been used to develop models for T_g [337, 339–341], thermal conductivity [342], dielectric constant [343], crystallization tendency [344] and bandgap [345].

The aforementioned works and a lot of others in the polymer informatics literature focus on homopolymers alone, and train over molecular descriptors without using any synthesis information. Such approaches can be reasoned by the widespread availability of homopolymer data in the databases [337]. For instance, PoLyInfo (accessed on January 31, 2022) lists T_g for 7996 homopolymers, 4135 composites, and 1156 blends. Upon filtering the characterization conditions to 'neat resin' material type and 'pellet' shape, T_g is available for only 161 homopolymers, 29 composites, and 18 blends. The stark difference in the number of data points highlights the lack of information about synthetic parameters, and non-standard experimental characterization approaches. Such differences can have a significant effect on the training of ML models, hence, necessitating the assembly of data sets with standardized synthesis protocols.

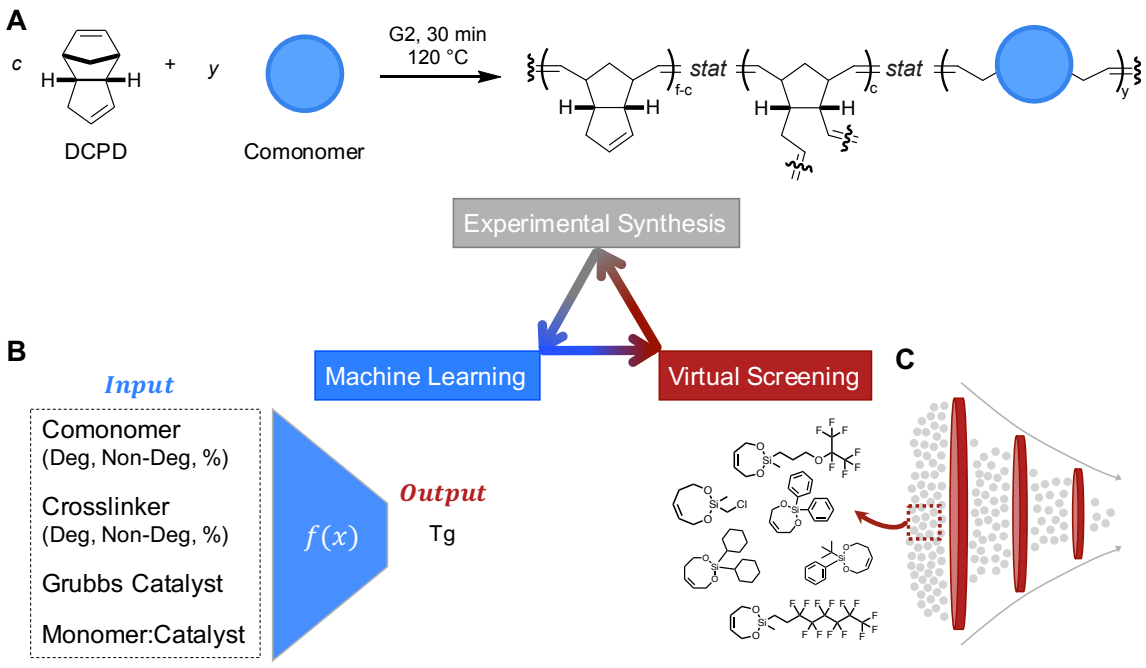


Figure 4-7: **Data-driven design of sustainable thermosets.** A. Experimental synthesis and validation of comonomer substituted DCPD polymers. B. ML framework to learn T_g from input features. C. Virtual screening of possible comonomer structures.

In this work, we developed a synthesis-learning-screening closed-loop framework to enable the discovery of new CCAs (Figure 4-7). Experimental synthesis of pDCPD derivatives and characterization of T_g for these polymers helped in acquiring a large, standardized data set. This data set was used to train a model ensemble, with varying model architectures, train-valid-test data sets, and hyperparameters, to predict T_g based on the comonomer, crosslinker molecules, and other experimental conditions. Using this model ensemble, we screened for a variety of molecular structures of cyclic alkene CCAs with different ring sizes containing silyl ether groups, and explored a range of comonomer and crosslinker compositions.

4.2.1 Data set

A data set, with information about chemicals and synthesis parameters, was assembled from existing works in the literature and in-house experiments [346–353]. A total of 101 data points, with different combinations of comonomer and crosslinker

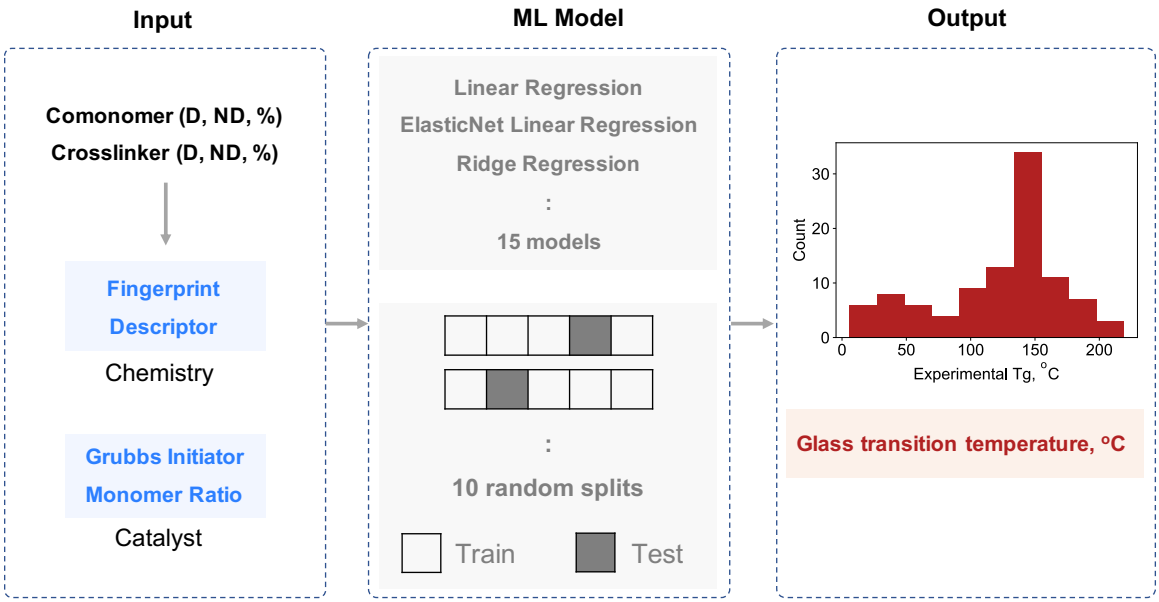


Figure 4-8: **Schematic of ML model.** Comonomer and crosslinker molecules, and other experimental features are used to train the model ensemble, consisting of different model architectures and sets of data, to predict T_g. The distribution of T_g values in the training data set is shown in the Output panel.

molecules, ratios of monomer to Grubbs initiator, and initiator type, along with their T_g values, measured after curing at 120 °C for 30 minutes, were collected.

The number of points in this data set falls in the low data regime for ML modeling. However, a key differentiator in this data set is the detailed information about synthesis parameters and standardized conditions of measuring T_g. This difference provides a unique opportunity for robust prediction of T_g, unlike the homopolymer-T_g data sets routinely used for training T_g models [337, 339–341].

4.2.2 Machine learning

We used different circular fingerprint representations of comonomer and crosslinker, and their respective concentrations in mol%, monomer to initiator ratio, and initiator type, as the input to the model ensemble (Figure 4-8). The mol% and monomer to initiator ratio were represented as continuous values, while the initiator type was represented using a one-hot encoding. To benchmark the molecular representations, we used physicochemical descriptors to train another set of models.

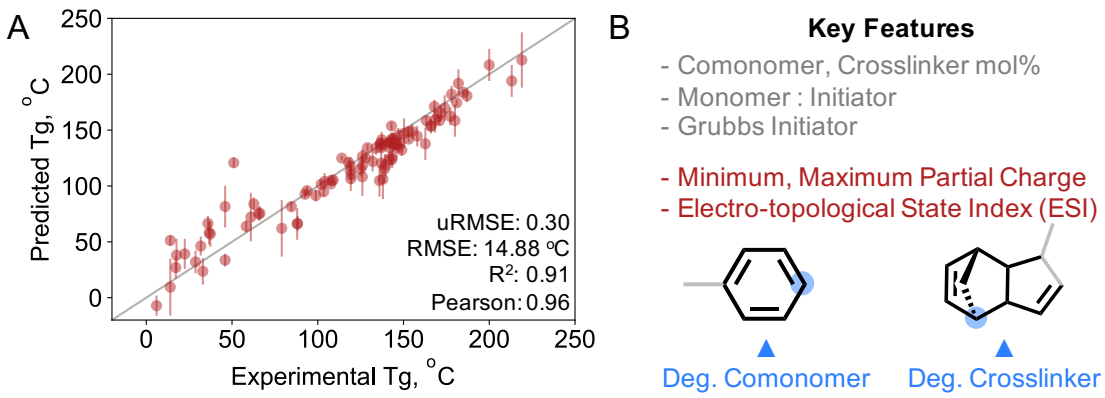


Figure 4-9: **Performance and attribution analysis of ML model.** A. Parity plot of predicted and experimental T_g values using the ML model ensemble. B. Key features affecting the T_g values identified using attribution analysis.

To address the limited data challenge, we trained an ensemble of 50 models, with 5 different hyperparameter-optimized model architectures, trained on 10 different parts of the training and validation data sets (Figure 4-9A). The top 5 architectures were chosen after analyzing a total of 12 architectures, based on linear models, random forest, gradient boosting, feed-forward networks and Gaussian process regression. The data set was split into 60:20:20 for training, validation and evaluation of the model ensemble. Our model could predict T_g for compositions in the held out test data set within 14.88 °C, and had a strong R^2 of 0.91, and a Pearson's correlation of 0.96.

Our strategy was to spread the learning across all models in the ensemble, such that we could capture the black-box function of mapping input features to the T_g, without overfitting to either a particular set of training data or to a particular functional form in the set of model architectures. This approach is in direct contrast to the usual convention of using cross-validation with a single model architecture, such as the works using only Gaussian process [339, 354], or convolutional neural networks [355, 356].

We used feature importance analysis for models in the ensemble, wherever possible, to identify key input features that impact the predicted T_g (Figure 4-9B). Consistently across both fingerprint and descriptor-based models, we noticed that the

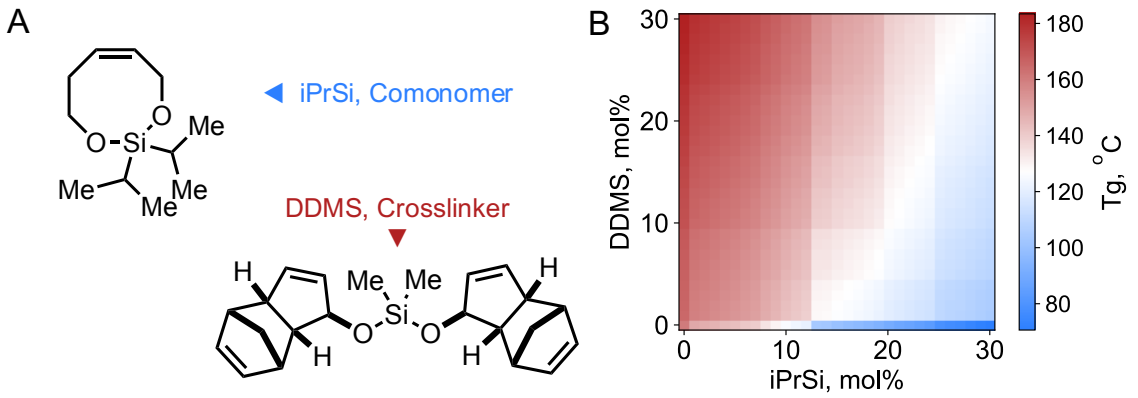


Figure 4-10: ***In silico* titration analysis with iPrSi and DDMS.** Variation of predicted T_g at different ratios of iPrSi and DDMS, while keeping all other experimental factors constant.

comonomer mol%, monomer to initiator ratio, and initiator type were the most important. Aryl substituents and norbornene-based crosslinkers were identified as the key substructures from the analysis of the fingerprint-based models, along with properties like number of aliphatic carbocycles, electrotopological state index (ESI), and partial charges, from the descriptor-based models.

4.2.3 *In silico* titration analysis and validation

To asses how well the model had learnt the property landscape, we varied the concentrations of iPrSi and DDMS across a grid of mol% values, and visualized the predicted T_g (Figure 4-10). iPrSi, a comonomer, copolymerizes with DCPD introducing cleavable bonds within the polymer network, and decreasing the network density, thus the T_g, while DDMS, a crosslinker, forms multiple links in the network, having an inverse effect. This contrasting characteristic was observed in the heatmap of T_g values. We experimentally synthesized and characterized selected combinations with {0%, 10%, 20%} iPrSi and {10%, 20%} DDMS. The experimentally observed T_g closely matched the predicted T_g values, hence successfully validating the model ensemble. In this experiment, we also discovered a novel composition of 10% iPrSi, 10% DDMS which had a T_g, 13 °C, greater than the previous best CCA-based thermoset, and closer to the T_g of pDCPD alone, 166.5 ± 5 °C [332].

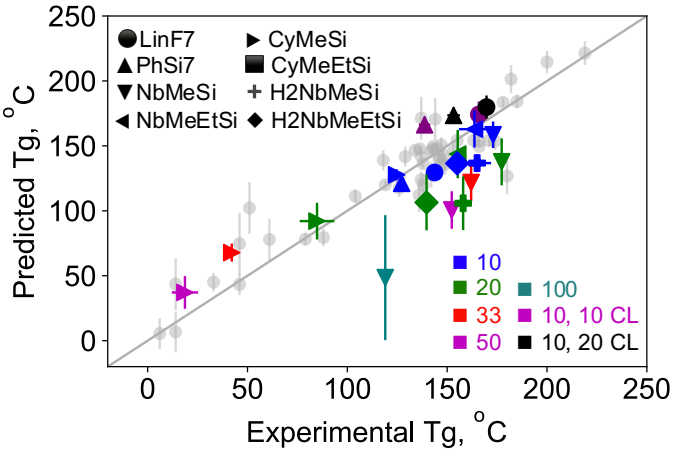


Figure 4-11: **Validation of predicted results.** Predicted and experimental T_g for the two screened molecules, LinF7 and PhSi7, and retrospective validation of other crosslinking comonomers, at different compositions.

4.2.4 Virtual screening and validation of new comonomers

With the success of the titration experiment, we attempted to expand the chemical space by screening possible substituents and comonomers that could be used to tune the T_g (Figure 4-11). We screened a library of commercially available dichlorosilanes, which could be used as CCAs or comonomers, and identified three molecules of interest - PhSi7, LinF7, and PFP7. We were able to synthesize and characterize T_g for PhSi7 and LinF7, noting that the predicted values matched the experimentally observed T_g. However, we were unable to synthesize a stable thermoset with PFP7. In addition to the T_g values, we experimentally confirmed that all new thermoset compositions were chemically deconstructable.

In another experiment, we retrospectively analyzed T_g for recently reported CCA-like strand-cleaving crosslinkers (SCCs) [336]. Although the model was not trained on this class of molecules, it was able to predict the T_g for all but one composition. This particular composition was a 100 mol% SCC without any DCPD, and not an additive alone. Understandably, the model was not able to predict the T_g and had a high uncertainty associated with it, demonstrating the low confidence attributed by the model ensemble. We believe that an active learning strategy with new classes of molecules would help generalize the model ensemble's prediction ability.

Chapter 5

Extended applications to small molecules

5.1 Design of organic photoelectronic molecules with desired properties

Deep generative models have been used to train on unlabeled data, sample novel molecules, but using them to design optimal molecules with desired properties remains a challenge [276, 357]. Recently, a new generative model framework based on conditional recurrent neural networks (cRNNs) was reported, and applied to the generation of drug-like molecules [75]. Unlike previous generative models that sample the latent space and discover novel molecules serendipitously, cRNNs translate desired property to string-like simplified molecular-input line-entry system (SMILES) representations.

The cRNN model enables the sampling novel of molecules conditioned on desired features, such as circular fingerprints, or desired properties. The model architecture combines a set of fully-connected neural network layers, followed by a recurrent neural network - long short-term memory (RNN-LSTM) decoder [105]. The fully-connected layers transform the input representation into a low-dimensional embedding, which is then used to set the initial state of the decoder. For sampling, fingerprint representa-

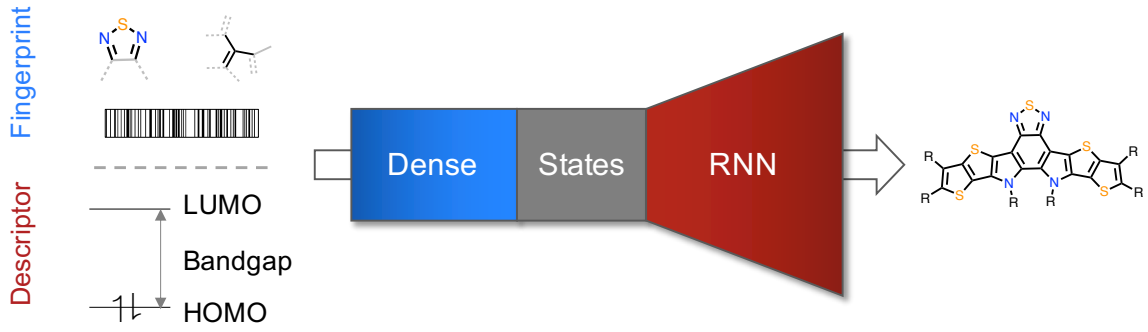


Figure 5-1: **Schematic of cRNN model.** The model translates chemical representations, either circular fingerprints or properties, into the SMILES-string for the molecule.

tions, or list of desired properties of molecules, used as-is or with added random noise, steer the generation of new molecules. Using a stronger supervision constrained by the properties, as compared to earlier approaches, the cRNN model attempts to constrain the chemical space to the desired properties [358–360]. Such a model could be used for inverse design of molecules with desired properties, such as drugs, metal-organic frameworks, or organic photoelectronic molecules (OPMs).

OPMs have a wide range of applications, from components of displays to solar cells [276, 361]. Although not chemically as diverse as drug-like molecules, OPMs have a vast chemical space with conjugated heterocycles and sizes in tens of heavy atoms. The desired electronic and optical properties for OPMs can be quantified using energies of the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), and the energy needed to transition from the HOMO to LUMO, called the optical gap. Using DFT simulations, the energy levels can be obtained with reasonable accuracy.

In this work, we trained cRNN models over fingerprints (FP) and DFT-calculated energies to generate new OPMs (Figure 5-1). Using a data set of OPMs, reported in the literature, and combinatorially generated molecules, we trained a large, unsupervised FP-based (FPB) model, and then used transfer learning (TL) to learn over molecules, with HOMO and optical gap values in the desired property range. We used the HOMO, LUMO and optical gap values, outside the desired properties range,

to train a physicochemical properties-based (PCB) model to translate the properties to molecules directly. The desired property range was HOMO between -7.5 and -6.5 eV, and optical gap between 2 and 3 eV, chosen based on HOMO -7 eV and optical gap 2.5 eV at which the solar flux has the highest intensity [362].

5.1.1 Data set

An unlabeled data set of OPMs was extracted by going through the literature and data available from the US patents and trademark office for the period 2001-19 [312, 363]. OPMs collected from literature and patents were fragmentized by disconnecting non-ring aryl single bonds in each molecule. For each set of fragments generated from their respective OPM, if one fragment occurs more than once in the OPM and is sandwiched between two different fragments, it is categorized as a functional group; otherwise, it is categorized as a capping group. If the fragment occurs only once, it is categorized as a core. Cores and functional groups were then combinatorically assembled to generate new molecules. This process was repeated once between the generated core-functional-group new molecules and capping groups to further generate new molecules. A total of 157,665 unlabelled molecules were assembled or generated in this process.

Based on the availability of similar experimental studies, heuristics, and random selection, DFT calculations were performed on a subset of the molecules to obtain HOMO, LUMO and optical gaps for the OPMs. Within the labeled data set, all molecules with HOMO and optical gap in the desired range were removed and marked as the 'seed' data set, and the remaining data set was used for training the TL and PCB models. The filtered data set had 13,616 molecules, while the seed data set had 1,129 molecules.

5.1.2 Machine learning

The FPB and TL models were trained on circular fingerprints of molecules, while the PCB model was trained on the property vectors (Figure 5-2). Non-canonical

A

Model	Validation Loss, CCE	Valid SMILES, %	Reconstruction, %	Unique Sampled Molecules, %	Novelty, %
FPB	0.025 ± 0.002	82.6 ± 0.3	20.7 ± 0.3	98.1 ± 0.4	67.4 ± 0.5
TL	0.217 ± 0.004	13.1 ± 0.3	0.0 ± 0.0	100.0 ± 0.0	98.5 ± 0.0
PCB	0.162 ± 0.004	42.4 ± 1.4	-	9.6 ± 0.6	91.3 ± 0.4

B

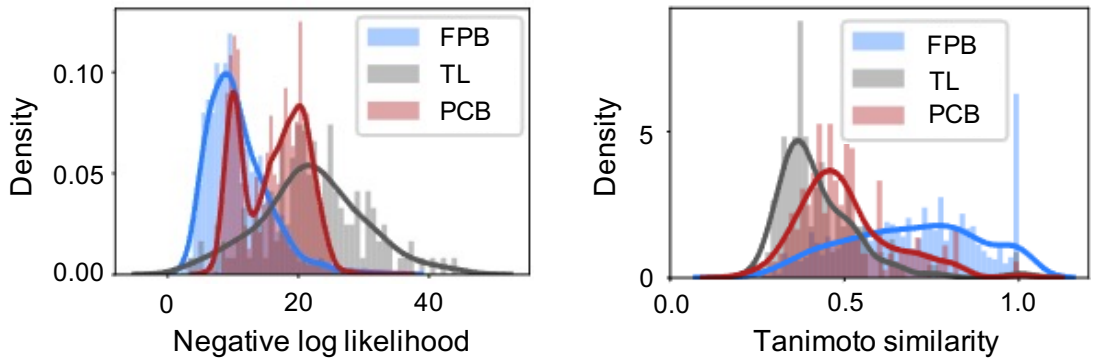


Figure 5-2: **Performance of cRNN models.** A. Different metrics for assessing the performance of FPB, TL and PCB models. B. Distributions of average negative log-likelihood and maximum Tanimoto similarity of the sampled molecules against the training data set.

variants of SMILES strings were used to augment the training data set. A 90:10 training: validation random split was used for the model training. The complete unlabeled data set, with the exception of the molecules in the labeled data set, was used to train the FPB model. The best performing FPB model had a categorical cross entropy (CCE) validation loss of 0.025. We re-trained this model using the filtered data set for 200 epochs to obtain the TL model. The TL model performed poorly compared to the FPB model given the significantly limited data for the labeled OPMs and had a CCE validation loss of 0.217. The PCB model, trained over the filtered data set had a CCE validation loss of 0.162.

We also used a variety of text-based, novelty and similarity metrics to assess the performance of these models. Valid SMILES % is the percentage of SMILES that can be processed using RDKit, reconstruction % is the percentage of molecules that were

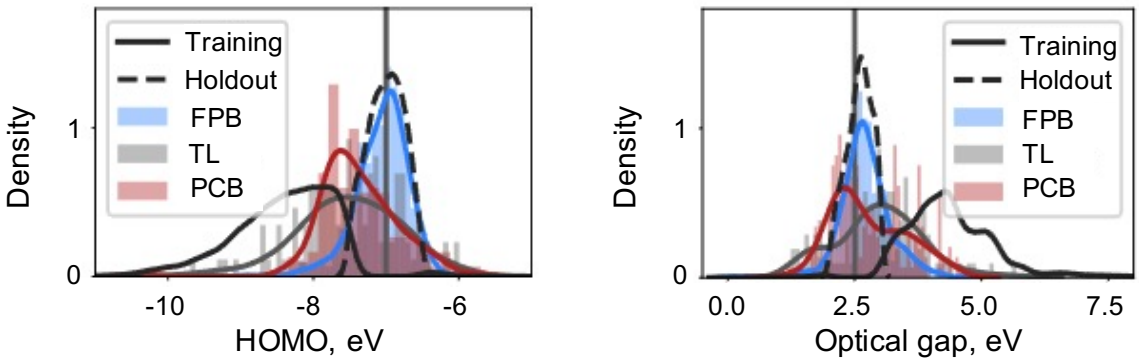


Figure 5-3: **Sampling of new molecules with desired properties.** Distributions of training, and held-out data sets for molecules used to seed FPB and TL models are shown, along with DFT-calculated HOMO and optical gap values for the molecules sampled using respective models. The vertical lines mark the desired properties of -7 eV HOMO, and 2.5 eV optical gap.

exactly the same as the seed molecules at inference time, unique sampled molecules % notes the percentage of unique molecules sampled across all the seeds, and novelty % notes how many new molecules not present in the training data set were sampled. The performance varied across the models with the FPB and PCB models generating a large number of valid SMILES, 82.6% and 42.4%, as compared to 13.1% for the TL model. The unique molecules generated across all the seeds was high in FPB and TL models as 98.1% and 100%, compared to 9.6% for PCB model, resulting from a possible mode collapse. In terms of novelty of molecules, TL and PCB models outperformed FPB model, with 98.5% and 91.3% novelty, respectively, compared to 67.4% for the FPB model.

The distributions of negative log-likelihood (NLL) and Tanimoto similarity indicate a trend similar to the aforementioned metrics across the models. The FPB model is noted to be good at reconstruction based on the lower NLL, while the TL and PCB models are good at generating novel molecules, as observed in the low maximum Tanimoto similarity of the molecules when compared to all of the training data set.

5.1.3 Conditional sampling

The pre-trained models were used to sample potential OPMs with desired properties (Figure 5-3). For the FPB and TL models, fingerprints of molecules in the seed data set were used. Randomly sampled property values within 0.5 eV of the desired HOMO and optical gap were used to seed the PCB model. We used DFT calculations to validate the properties of these sampled molecules, and observed that all the models generated molecules that were close to the desired HOMO of -7 eV, with HOMOs of -7.07 $\pm$ 0.41, -7.50 $\pm$ 0.70, and -7.41 $\pm$ 0.67 eV, as well as the desired optical gap of 2.5 eV, with gaps of 2.72 $\pm$ 0.49, 2.97 $\pm$ 0.85, and 2.78 $\pm$ 0.73 eV, for FPB, TL, and PCB models, respectively.

In the sampling experiment, we also noted that HOMO and optical gap values are not only closer to the desired property values, but also significantly different from the mean of the training data set. This difference indicates that the models were able to learn the general chemical space of OPMs, and help in sampling molecules in the desired property range, despite not having been trained explicitly over them.

5.1.4 Benchmarking against graph-based genetic algorithm

A graph-based genetic algorithm (GB-GA) approach was used to benchmark the generation of OPMs using cRNN models (Figure 5-4) [364]. 100 random molecules were chosen from the unlabeled data set to seed the GA, and default settings were used for all mutations. The objective function in the GA was set to minimize the mean squared error of predicted and desired HOMO and optical gap values. The predictions for HOMO and optical gap were obtained using a pre-trained graph-convolutional model trained on OPMs in the labeled data set, with validation mean absolute error of 0.24 eV for the sum of the two properties [365].

In this experiment, we noted that both approaches produced a significant number of unrealistic molecules, and only a small number of realistic OPMs. Two molecules generated using cRNN, first and fifth in the Figure 5-4 were noted to be realistic, while none in the GB-GA set were found to be realistic. The unrealistic molecules

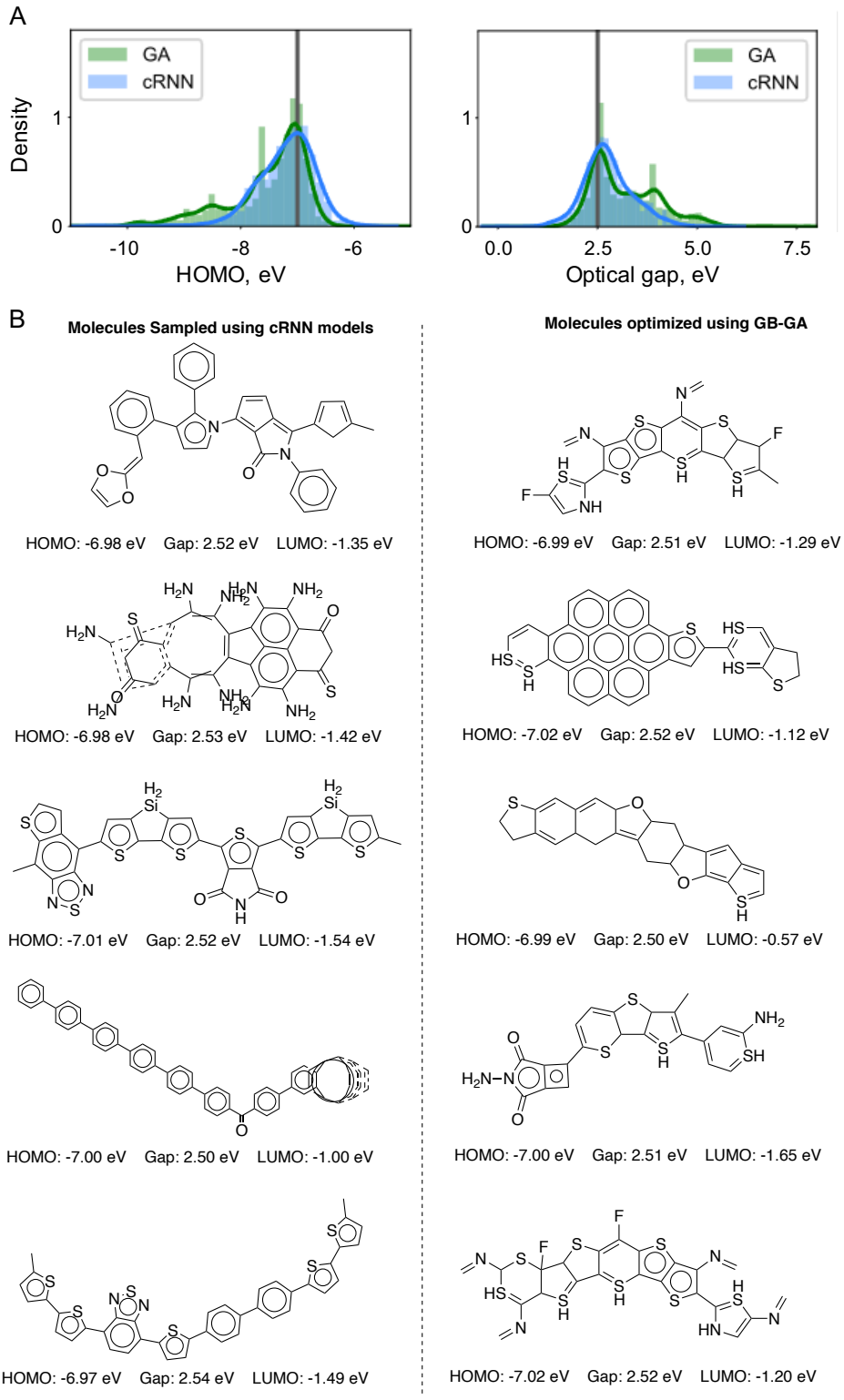


Figure 5-4: **Benchmarking cRNN against GB-GA.** A. Distribution of HOMO and optical gap of molecules, generated using cRNN and GB-GA. B. Visualization of 5 randomly selected molecules, along with the properties noted as text.

generated using the GB-GA can be reasoned by the reliance on the curated set of rules that the approach uses to mutate the molecules. The lack of chemical context in the molecule generation process affects the sampling. For the cRNNs, we noted that the decoding of SMILES played a role in the unfeasible OPMs. We believe that increasing the amount of training data and the size of the models will help improve the generation using the cRNN models.

5.2 Inverse design of molecules from shapes with desired chemistry

Inverse design of materials with specific properties could help accelerate development in a wide range of areas, from catalysis to therapeutics to photonics [357, 366–369]. Deep generative models, such as variational autoencoders (VAEs), have been used to design organic photovoltaics [276, 370], porous materials [371, 372], inorganic solid state materials [373–375], and several other types of materials. However, such models usually require a large amount of training data, and often lead to generation of a large number of nonsensical molecules. Recently, a conditional generative model architecture based on recurrent neural networks (cRNNs) was introduced. cRNNs map property or chemical feature vectors directly to molecules, constraining the chemical space by stronger supervision [75, 76].

Individual molecules exist in a number of 3-dimensional structures, or conformers. Ensemble and individual properties of conformers, and the 3D structures themselves, are routinely used to inform physicochemical and biological processes, such as reaction mechanisms [376], electrocatalysis [377], and protein-ligand interactions [378, 379].

A significant amount of effort has been spent in developing physics and heuristic-driven methods to predict conformers, given a molecule [380–383]. Similarly, a number of machine learning models, such as GeoMol [384] and others [385–388], have also been recently developed to address the forward conformer generation. In terms of cardinality, aforementioned approaches belong to a class of one-to-many relationships,

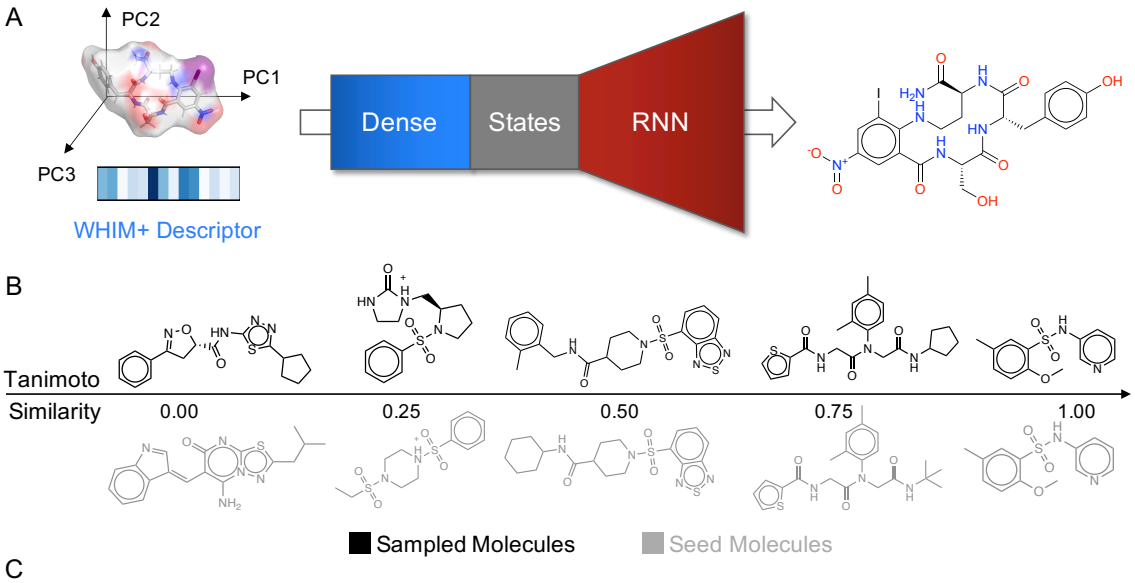
mapping a single molecule to multiple conformers.

Several works have tried to do the inverse, attempting to map 3-dimensional shape and property information, to a molecule, proposing a many-to-one model [389]. De Fabritiis and co-workers used 3D representations, such as voxels of ligands [390] and target protein pockets [391], with 3D-convolutional neural network-based VAEs, to generate molecules. Koes and co-workers used a similar approach with 3D representations, mapping the spatial atomic density to individual atom types, followed by generation of the SMILES representation of the molecule [392, 393]. cRNN models have been used to train over the eigenvalues of the Coulomb matrix as a descriptor for the protein pocket to design ligands [394]. Recently, both voxels and molecule graphs have been used to train multi-modal models, such as DEVELOP and DeLinker [395, 396]. However, it maybe noted that most of the models mentioned here use 3D representations - voxels, molecular graphs or spatial atomic density - which might be passing information about the atomic connectivities to the models. We believe that such representations might limit the model in learning a small chemical space, which in turn only partially learns the general mapping function of shape to molecule.

To address this challenge, we developed a model framework, Shape2Mol (S2M), which translates conformers, represented using weighted holistic invariant molecular (WHIM) and a few other global physicochemical descriptors, to SMILES strings (Figure 5-5A). We used this model to generate rigid versions of small molecules, and quantified them using a variety of metrics. In addition, we generated potential binders for three different proteins with different levels of docking difficulty - DRD2 kinase (easy), F2 protease (medium), and ESR2 nuclear receptor (hard); and small molecule analogues of peptides using their predicted 3D structures.

5.2.1 Data set

We used conformers for drug-like molecules from the GEOM data set [397]. Briefly, the GEOM data set was assembled by running semi-empirical DFT calculations and high quality sampling using the CREST program [381, 398]. From more than 300,000 molecules in GEOM, we randomly sampled 30,000 molecules for training, and another



C

# Molecules	# Conformers	CCE Loss		Metrics for generated molecules					
		Train	Valid	% Valid	% Reconstruction	% Unique	Tanimoto Similarity		
							Mean	Minimum	Maximum
10,000	1	8.00E-06	0.107	0.00	-	-	-	-	-
	3	1.00E-06	0.058	100.00	0.00	0.00	0.24	0.07	0.46
	10	1.60E-05	0.080	85.80	0.10	85.80	0.42	0.09	1.00
30,000	1	4.00E-06	0.041	91.40	0.20	91.40	0.43	0.07	1.00
	3	1.00E-06	0.049	100.00	0.00	0.00	0.23	0.06	0.50
	10	5.07E-04	0.041	91.10	0.30	91.10	0.49	0.07	1.00

Figure 5-5: **Overview of Shape2Mol model.** A. Schematic of Shape2Mol model, trained with WHIM and physicochemical descriptors of conformers as input, and SMILES representations of molecules as output. B. Visualization of molecules, generated using cRNN model, with different Tanimoto similarities with the seed molecules. C. CCE loss and other metrics, as described in Section 5.1.2, for models trained on 10,000 and 30,000 different molecules, with {1, 3, 10} conformers. The shaded model with 30,000 molecules and 10 conformers was used for further analysis.

1,000 molecules for the held-out test data set. One or more conformers for a single molecule was used for model training and evaluation.

5.2.2 Machine learning

We trained cRNN models over conformers represented using a feature vector of WHIM [71, 399] and physicochemical descriptors to generate SMILES strings (Figure 5-5C). All the physicochemical descriptors used here are global properties without any explicit connectivity information, such as topological polar surface area [400], and electrotopological state index [401]. For each molecule, five non-canonical SMILES were

generated by random enumeration of the starting atom in RDKit, giving a total of six training data points per molecule. The training data set was split 80:20 for training and validation of the model, and the held-out test data set was used to evaluate the model post-training. Default hyperparameters were used for training all the models.

Multiple cRNN models were trained using different combinations of number of molecules and conformers per molecule. We observed that increasing both the number of molecules and the conformers helped in improving the performance of the model on the test data set.

5.2.3 Comparative Analysis of Generated Molecules

We hypothesized that sampling using the model might lead to more rigid molecules than the seed molecule, while preserving a similar spatial arrangement. To test this hypothesis, we carried out the sampling in two parts - using the lowest energy conformer, and using a random high energy conformer. In case of the lowest energy conformer, we sampled using descriptors of 1000 molecules from a held-out test data set. For sampling with the high energy conformer, we filtered the molecules from the test data set, to those having more than hundred unique conformers, and then choosing a conformer with energy greater than 5 kcal/mol relative to the most stable conformer. Visual inspection of the sampled molecules provided some evidence that the model uses classical medicinal chemistry tricks such as ring closing, and converting hexane to benzene rings (Figure 5-5B).

To rigorously assess the sampled molecules, we used a variety of metrics, ranging from cheminformatic descriptors, information obtained from conformation generation using CREST, to distances between descriptors.

At first, we calculated the number of rotatable bonds in the seed and generated molecules (Figure 5-6). We observed that when the lowest energy conformer was used, the distribution of the number of rotatable bonds was similar in both seed and generated molecules, with a mean of 5.00 and 4.99 rotatable bonds, respectively. For sampling with high energy conformers, the distribution obtained for the generated molecules had a lower mean 6.20 as compared to 6.78 for the seed molecules. Another

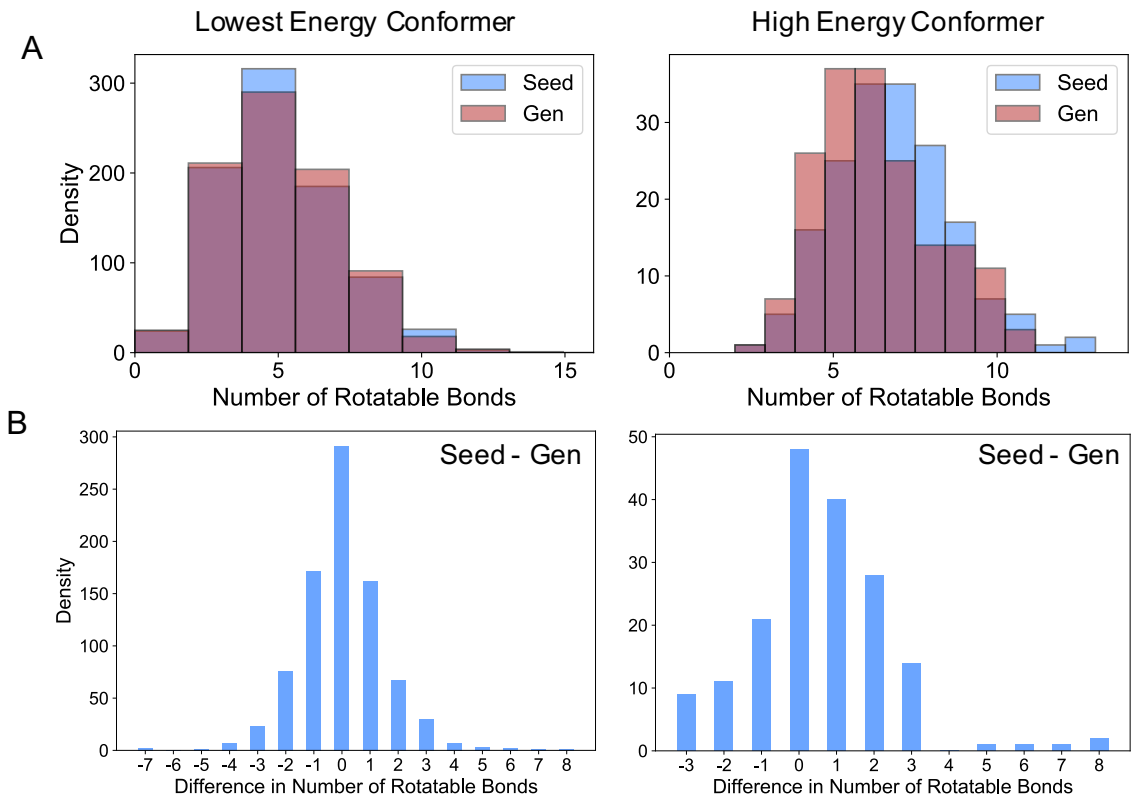


Figure 5-6: **Comparison of rotatable bonds in generated molecules.** A. Distributions of number of rotatable bonds in generated molecules, when S2M is seeded with lowest and a random high energy conformer, as compared to the seed molecules. B. Difference of number of rotatable bonds between seed and generated molecules, where >0: seed has more, <0: seed has less, =0: same number of rotatable bonds. Results from generation using lowest energy conformer and a random high energy conformer are shown in left and right columns.

perspective for the rotatable bonds was obtained by calculating the difference between the number of rotatable bonds in the seed and generated molecules. For the sampling using the lowest energy conformers, the bar plot was nearly Gaussian, indicating that the sampled molecules were indifferent to the rotatable bonds, thereby acting as a positive control. However, in sampling using high energy conformers, we observed that half of the sampled molecules had lower rotatable bonds than the seed molecule.

The SC RDKit score, as described in [396], was used to compare the exact conformers of seed and generated molecules (Figure 5-7A). For both sampling approaches, all results were reported for the conformer of the sampled molecule with the highest SC RDKit score against the seed conformer. The distributions for the scores for the

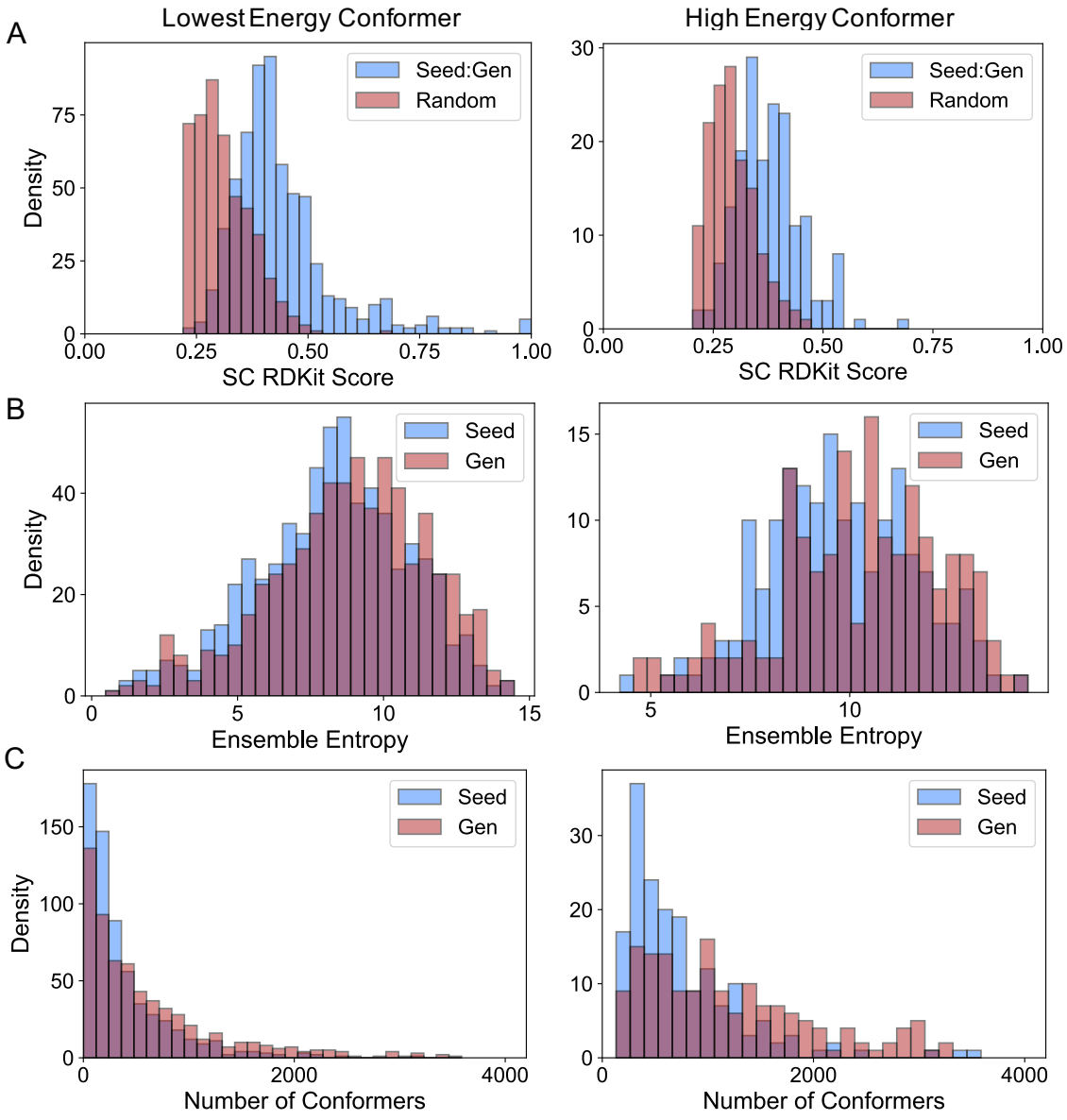


Figure 5-7: **Conformer comparison using similarity, entropy metrics, and count.** A. Distribution of similarity, calculated as SC RDKit score, of generated conformer with the seed conformer, compared to random conformers. B. Distribution of ensemble entropy of the generated conformers, compared to seed conformers. C. Distribution of number of unique conformers of seed and generated molecules. Results from generation using lowest energy conformer and a random high energy conformer are shown in left and right columns.

seed and generated conformers, and two random conformers, were different for both sampling approaches, with the model having a difficult time in sampling molecules when seeded with high energy conformer descriptors. Specifically, only 7% of the molecules sampled using the high energy conformer had a score greater than 0.5, as compared to 19% of the molecules when sampled using the lowest energy conformer.

The distribution of ensemble entropy was similar for seed and generated conformers, for sampling using lowest and high energy conformers (Figure 5-7B). The difference in the spread of the distribution, especially the wider spread for the lowest energy conformers, as compared to the narrow distribution of high energy conformers, may be attributed to the flexibility of the generated molecules.

The distributions of the number of conformers for seed and sampled molecules was different for both sampling approaches (Figure 5-7C). When sampling using the lowest energy conformer, both seed and sampled molecules had a right skewed distribution. For the high energy conformers, the seed molecules followed a similar distribution, as the lowest energy conformers, while the sampled molecules had a wide spread and more uniform distribution. This distribution may be attributed to the sampling of varied high energy spatial configurations.

5.2.4 Docking of cRNN generated molecules

We used the model to generate new molecules from docked poses of molecules bound to three different proteins. The three proteins - DRD2 kinase (easy), F2 protease (medium), and ESR2 nuclear receptor (hard) - were selected from the benchmarks presented in [402], noting proteins with a range of functions in the order of difficulty of being able to dock molecules. For each of the proteins, 1000 molecules with varying binding affinities or docking energies were randomly sampled from the DockString data set [402]. Conformers of these molecules were seeded to the pre-trained cRNN model to generate new molecules. These molecules were docked to their respective target proteins using the same approach that was used to obtain the seeds.

The models were able to generate molecules with low Tanimoto similarity to the seed molecules, while still having low docking energy or high binding affinity (Figure

5-8). Visualizing the docked poses of the seed and generated molecules together, we observed that the interactions by the molecules with the target proteins were similar in both cases, thereby validating our approach of exploring the local chemical space while preserving the shape of the molecules.

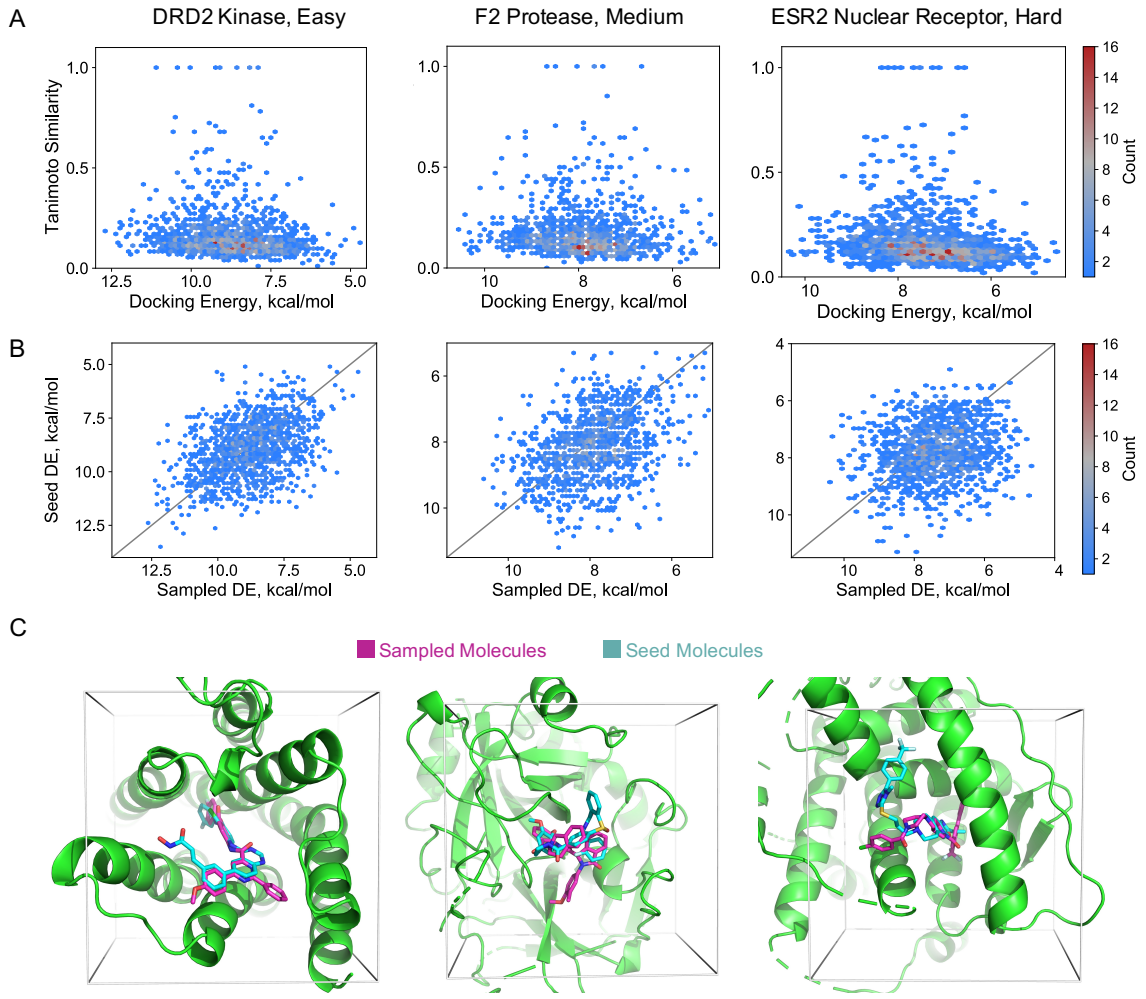


Figure 5-8: **Docking generated molecules against DRD2 kinase, F2 protease and ESR2 nuclear receptor proteins.** A. Docking energy and Tanimoto similarity of generated molecules for 3 proteins. B. Comparison of docking energies of seed and generated molecules. Data points above the $y = x$ line have higher binding affinity or lower docking energy for generated molecule, as compared to seed molecule, and vice-versa. C. Overlay of docked poses of generated molecules with the best docking energies, and corresponding seed molecules, with respective target proteins.

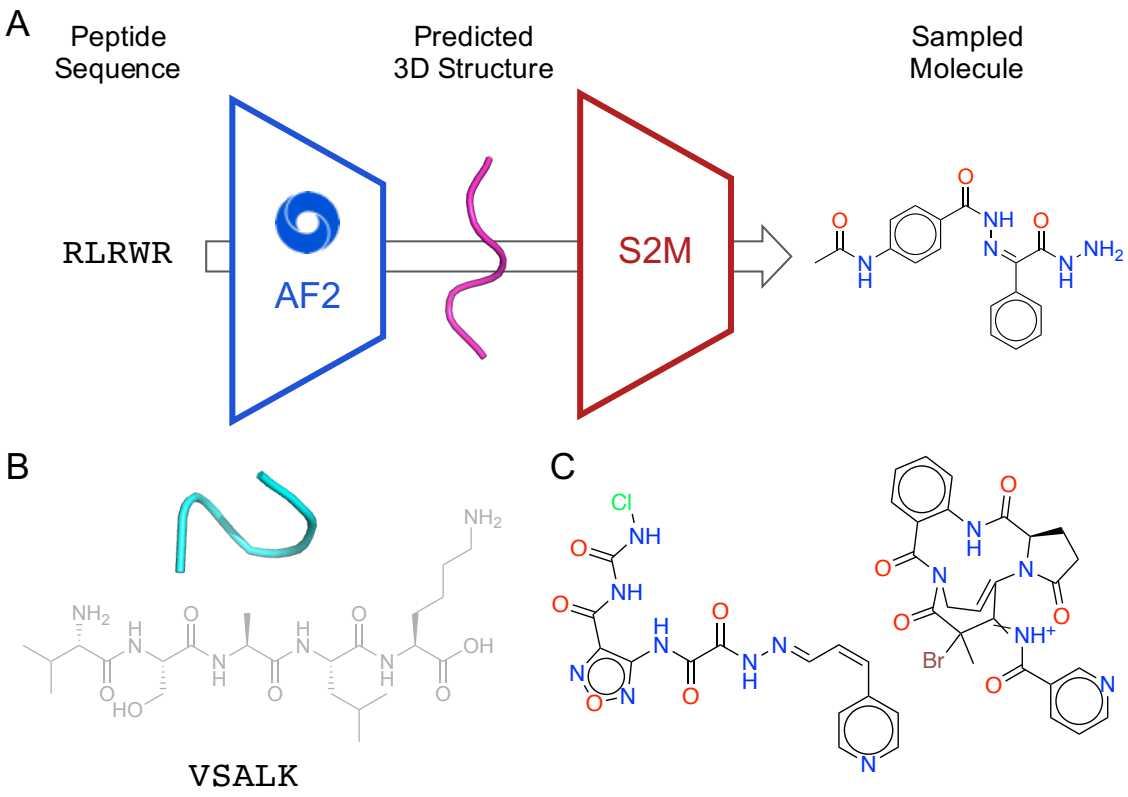


Figure 5-9: **Small molecule analogues of peptides.** A. Schematic of using AlphaFold2 to generate 3D structures from peptide sequences, and seeding them in S2M model to generate small molecules. B. 3D and 2D structures for peptide sequence - VSALK. C. Molecules generated, when seeded with 3D structure of VSALK.

5.2.5 Generation of small molecule analogues of peptides

We generated small molecules with conformers similar to predicted 3D structures of peptides (Figure 5-9). Peptide sequences with all-natural amino acids and lengths less than 8 residues were selected from the data set of cell-penetrating peptides [64]. The 3D structures of peptides were predicted using AlphaFold2 [63]. We used WHIM descriptors calculated for the top-ranked 3D structures as-is and with added random noise to seed the cRNN models. 2-4 valid molecules were sampled for each peptide. Tanimoto similarity of all generated molecules with corresponding peptides is within 0.15, with most of the molecules having a similarity of 0.1 or less. Using this approach, peptides and other macromolecule 3D structures can be translated into small molecule analogues with similar shapes.

Chapter 6

Conclusion

Our work presented a computational framework to represent, compute similarity for, and learn over sequence-defined macromolecules with arbitrary monomer/linkage chemistry and topology. We applied this approach to a wide range of biological and artificial macromolecules, as well as small molecules, developing interpretable models and achieving state-of-the-art or comparable results. Using these models, we virtually screened, designed *de novo*, or optimized, and experimentally validated materials with desired properties.

For biological macromolecules, our work involved optimization of functionality and synthetic accessibility of peptides, peptide-nucleic acids, and glycans. Quantitative prediction of cell-penetrating efficacy of mini-proteins and peptides enabled optimization of gene therapy delivery, opening doors to a number of other therapeutic modalities, which could be delivered using a similar platform. Unsupervised and supervised machine learning and dynamic time warping analysis helped in discovering sub-families of peptidomimetic binders, and resulted in design of new binders. Peptide vaccines for cancer were designed by machine learning of human degrons, and effectively harnessing the antigen processing and presentation by the ubiquitin-proteasome degradation system. Prediction and optimization of synthetic accessibility of peptides and peptide nucleic acids improved flow synthesis. Applications to non-linear biomacromolecules resulted in state-of-the-art or comparable results for a number of tasks, and discovery of novel high-affinity, lectin-specific glycans.

We developed robust prediction and uncertainty estimation approaches to learn over artificial macromolecules in the low-data regime. Using a physics-informed model, we were able to map the effect of secondary substituents in oligoethylene glycol-based Li-battery electrolytes to conductivity and viscosity, screen and successfully validate new substituted-electrolytes. We discovered and validated a novel comonomer: crosslinker composition, and another new comonomer, for chemically-deconstructable thermosets, with glass transition temperature comparable to commonly used polydicyclopentadiene thermosets.

Extending our work to small molecules, we used conditional generative models to design organic photoelectronic molecules with specific properties. Modifying the cRNN model architecture, we developed Shape2Mol to inversely design molecules from shape with desired chemistry. With this model, we generated molecules that docked better than seeds in target proteins, and small molecule analogues for peptides from their predicted 3D structures.

We believe our work serves as a foundation for further work on sequence-defined macromolecules. In the future, multi-objective optimization against different properties, such as functionality, synthetic accessibility, cost and carbon footprint, could be used to design better and more sustainable materials. We envision that our work will ultimately lead to acceleration of development of macromolecules across several domains.

Bibliography

[1] Paul C Hiemenz and Timothy P Lodge. *Polymer chemistry*. CRC press, 2007. ISBN 1420018272.

[2] Arthur Cayley. A theorem on trees. *Quart. J. Math.*, 23:376–378, 1889.

[3] Jeffries Wyman and Stanley J Gill. *Binding and linkage: functional chemistry of biological macromolecules*. University Science Books, 1990. ISBN 0935702563.

[4] Charles R Cantor and Paul R Schimmel. *Biophysical chemistry: Part III: the behavior of biological macromolecules*. Macmillan, 1980.

[5] Sebastian Pomplun, Zachary P Gates, Genwei Zhang, Anthony J Quartararo, and Bradley L Pentelute. Discovery of nucleic acid binding molecules from combinatorial biohybrid nucleobase peptide libraries. *Journal of the American Chemical Society*, 142(46):19642–19651, 2020.

[6] Tathagata Mondal, Maria Nerantzaki, Kevin Flesch, Capucine Loth, Mounir Maaloum, Yidan Cong, Sergei S Sheiko, and Jean-François Lutz. Large sequence-defined supramolecules obtained by the dna-guided assembly of biohybrid poly (phosphodiester) s. *Macromolecules*, 54(7):3423–3429, 2021.

[7] Robert G Spiro. Glycoproteins. *Advances in protein chemistry*, 27:349–467, 1973.

[8] RD Marshall. Glycoproteins. *Annual review of biochemistry*, 41(1):673–702, 1972.

[9] Eric Chevet, Maria Antonietta De Matteis, Eeva-Liisa Eskelinen, and Hesso Farhan. Rna, a new member in the glycan-club that gets exposed at the cell surface. *Traffic*, 22(10):362–363, 2021.

[10] Ryan A Flynn, Kayvon Pedram, Stacy A Malaker, Pedro J Batista, Benjamin AH Smith, Alex G Johnson, Benson M George, Karim Majzoub, Peter W Villalta, Jan E Carette, et al. Small rnas are modified with n-glycans and displayed on the surface of living cells. *Cell*, 184(12):3109–3124, 2021.

[11] Peter E Nielsen. Applications of peptide nucleic acids. *Current opinion in biotechnology*, 10(1):71–75, 1999.

[12] Peter E Nielsen. Peptide nucleic acids as therapeutic agents. *Current opinion in structural biology*, 9(3):353–357, 1999.

[13] Adrianne M Rosales, Rachel A Segalman, and Ronald N Zuckermann. Polypeptoids: a model system to study the effect of monomer sequence on polymer properties and self-assembly. *Soft Matter*, 9(35):8400–8414, 2013.

[14] Jean-Francois Lutz, Jean-Marie Lehn, E W Meijer, and Krzysztof Matyjaszewski. From precision polymers to complex materials and systems. *Nature Reviews Materials*, 1(5):1–14, 2016. ISSN 2058-8437.

[15] Jean-François Lutz, Makoto Ouchi, David R Liu, and Mitsuo Sawamoto. Sequence-controlled polymers. *Science*, 341(6146), 2013.

[16] Matteo Romio, Lucca Trachsel, Giulia Morgese, Shivaprakash N. Ramakrishna, Nicholas D. Spencer, and Edmondo M. Benetti. Topological Polymer Chemistry Enters Materials Science: Expanding the Applicability of Cyclic Polymers. *ACS Macro Letters*, 9(7):1024–1033, 2020. ISSN 21611653. doi: 10.1021/acsmacrolett.0c00358.

[17] Alfred J Crosby and Jong-Young Lee. Polymer nanocomposites: the "nano" effect on mechanical properties. *Polymer reviews*, 47(2):217–229, 2007. ISSN 1558-3724.

[18] Andrew J Boydston, Jianxun Cui, Chang-Uk Lee, Brock E Lynde, and Cody A Schilling. 100th Anniversary of Macromolecular Science Viewpoint: Integrating Chemistry and Engineering to Enable Additive Manufacturing with High-Performance Polymers. *ACS Macro Letters*, 9 (8):1119–1129, aug 2020. doi: 10.1021/acsmacrolett.0c00390. URL `https://doi.org/10.1021/acsmacrolett.0c00390`.

[19] Stefan Cichosz, Anna Masek, and Marian Zaborski. Polymer-based sensors: A review. *Polymer testing*, 67:342–348, 2018. ISSN 0142-9418.

[20] Chase B Thompson and LaShanda T J Korley. 100th Anniversary of Macromolecular Science Viewpoint: Engineering Supramolecular Materials for Responsive Applications—Design and Functionality. *ACS Macro Letters*, 9(9):1198–1216, sep 2020. doi: 10.1021/acsmacrolett.0c00418. URL `https://doi.org/10.1021/acsmacrolett.0c00418`.

[21] Huanli Sun and Zhiyuan Zhong. 100th anniversary of macromolecular science viewpoint: Biological stimuli-sensitive polymer prodrugs and nanoparticles for tumor-specific drug delivery. *ACS Macro Letters*, 9(9):1292–1302, 2020. ISSN 21611653. doi: 10.1021/acsmacrolett.0c00488.

[22] Jean François Lutz. Coding Macromolecules: Inputting Information in Polymers Using Monomer-Based Alphabets. *Macromolecules*, 48(14):4759–4767, 2015. ISSN 15205835. doi: 10.1021/acs.macromol.5b00890.

[23] Charles Yoon-hyung Cho, Edmund J Moran, JC Stephans, SP Fodor, Cynthia L Adams, Arathi Sundaram, Jeffrey W Jacobs, Peter G Schultz, et al. An unnatural biopolymer. *Science*, 261(5126):1303–1305, 1993.

[24] Michael J Soth and James S Nowick. Unnatural oligomers and unnatural oligomer libraries. *Current opinion in chemical biology*, 1(1):120–129, 1997.

[25] Philipp M Cromm, Jochen Spiegel, and Tom N Grossmann. Hydrocarbon stapled peptides as modulators of biological function. *ACS chemical biology*, 10 (6):1362–1375, 2015.

[26] Scott G Gaynor, Shane Edelman, and Krzysztof Matyjaszewski. Synthesis of branched and hyperbranched polystyrenes. *Macromolecules*, 29(3):1079–1081, 1996.

[27] Haifeng Gao and Krzysztof Matyjaszewski. Synthesis of star polymers by a combination of atrp and the "click" coupling method. *Macromolecules*, 39(15): 4960–4965, 2006.

[28] Nikos Hadjichristidis. Synthesis of miktoarm star (μ-star) polymers. *Journal of Polymer Science Part A: Polymer Chemistry*, 37(7):857–871, 1999.

[29] Jeremiah A Johnson, Ying Y Lu, Alan O Burts, Yeon-Hee Lim, MG Finn, Jeffrey T Koberstein, Nicholas J Turro, David A Tirrell, and Robert H Grubbs. Core-clickable peg-branch-azide bivalent-bottle-brush polymers by romp: grafting-through and clicking-to. *Journal of the American Chemical Society*, 133(3):559–566, 2011.

[30] Tianhong Zhang, Hongli Li, Hualin Xi, Robert V Stanton, and Sergio H Rotstein. Helm: a hierarchical notation language for complex biomolecule structure representation, 2012.

[31] Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):1–34, 2015.

[32] Axel Drefahl. Curlysmiles: a chemical language to customize and annotate encodings of molecular and nanodevice structures. *Journal of cheminformatics*, 3(1):1–7, 2011.

[33] Tzyy-Shyang Lin, Connor W Coley, Hidenobu Mochigase, Haley K Beech, Wencong Wang, Zi Wang, Eliot Woods, Stephen L Craig, Jeremiah A Johnson, Julia A Kalow, et al. Bigsmiles: a structurally-based line notation for describing macromolecules. *ACS central science*, 5(9):1523–1531, 2019.

[34] Daniel Bojar, Rani K Powers, Diogo M Camacho, and James J Collins. Deep-learning resources for studying glycan-mediated host-microbe interactions. *Cell Host & Microbe*, 29(1):132–144, 2021.

[35] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[36] Chiho Kim, Anand Chandrasekaran, Tran Doan Huan, Deya Das, and Rampi Ramprasad. Polymer genome: a data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C*, 122(31):17575–17585, 2018.

[37] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215 (3):403–410, 1990.

[38] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[39] Grzegorz M Boratyn, Jean Thierry-Mieg, Danielle Thierry-Mieg, Ben Busby, and Thomas L Madden. Magic-blast, an accurate rna-seq aligner for long and short reads. *BMC bioinformatics*, 20(1):1–19, 2019.

[40] Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[41] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[42] Sean R Eddy. Where did the blosum62 alignment score matrix come from? *Nature biotechnology*, 22(8):1035–1036, 2004.

[43] Kaspar Riesen and Horst Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision computing*, 27(7):950–959, 2009.

[44] Zhenjie Zhang, Marios Hadjieleftheriou, Beng Chin Ooi, and Divesh Srivastava. Bed-tree: an all-purpose index structure for string similarity search based on edit distance. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 915–926, 2010.

[45] Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575. World Scientific, 2001.

[46] Tommi Jaakkola, Mark Diekhans, and David Haussler. A discriminative framework for detecting remote protein homologies. *Journal of computational biology*, 7(1-2):95–114, 2000.

[47] Maxwell L Bileschi, David Belanger, Drew H Bryant, Theo Sanderson, Brandon Carter, D Sculley, Mark L DePristo, and Lucy J Colwell. Using deep learning to annotate the protein universe. biorxiv, 2019.

[48] Seokjun Seo, Minsik Oh, Youngjune Park, and Sun Kim. Deepfam: deep learning based alignment-free method for protein family modeling and prediction. *Bioinformatics*, 34(13):i254–i262, 2018.

[49] Limin Li, Wai-Ki Ching, Takako Yamaguchi, and Kiyoko F Aoki-Kinoshita. A weighted q-gram method for glycan structure classification. *BMC bioinformatics*, 11(1):1–6, 2010.

[50] Kiyoko F Aoki, Atsuko Yamaguchi, Yasushi Okuno, Tatsuya Akutsu, Nobuhisa Ueda, Minoru Kanehisa, and Hiroshi Mamitsuka. Efficient tree-matching methods for accurate carbohydrate database queries. *Genome informatics*, 14:134–143, 2003.

[51] Masae Hosoda, Yukie Akune, and Kiyoko F Aoki-Kinoshita. Development and application of an algorithm to compute weighted multiple glycan alignments. *Bioinformatics*, 33(9):1317–1323, 2017.

[52] Lachlan Coff, Jeffrey Chan, Paul A Ramsland, and Andrew J Guy. Identifying glycan motifs using a novel subtree mining approach. *BMC bioinformatics*, 21 (1):1–18, 2020.

[53] Yoshihiro Yamanishi, Francis Bach, and Jean-Philippe Vert. Glycan classification with tree kernels. *Bioinformatics*, 23(10):1211–1216, 2007.

[54] David B Blumenthal and Johann Gamper. On the exact computation of the graph edit distance. *Pattern Recognition Letters*, 134:46–57, 2020.

[55] Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray, and Bastian Rieck. Graph kernels: State-of-the-art and future challenges. *arXiv preprint arXiv:2011.03854*, 2020.

[56] Nils M Kriege, Fredrik D Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 5(1):1–42, 2020.

[57] Mahito Sugiyama, M Elisabetta Ghisu, Felipe Llinares-López, and Karsten Borgwardt. graphkernels: R and python packages for graph comparison. *Bioinformatics*, 34(3):530–532, 2018.

[58] Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis, and Michalis Vazirgiannis. Grakel: A graph kernel library in python. *J. Mach. Learn. Res.*, 21:54–1, 2020.

[59] Jon Paul Janet and Heather J Kulik. *Machine Learning in Chemistry*. American Chemical Society, 2020.

[60] Adam C Mater and Michelle L Coote. Deep learning in chemistry. *Journal of chemical information and modeling*, 59(6):2545–2559, 2019.

[61] Nongnuch Artrith, Keith T Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh. Best practices in machine learning for chemistry. *Nature Chemistry*, 13(6):505–508, 2021.

[62] Raphael JL Townshend, Stephan Eismann, Andrew M Watkins, Ramya Rangan, Maria Karelina, Rhiju Das, and Ron O Dror. Geometric deep learning of rna structure. *Science*, 373(6558):1047–1051, 2021.

[63] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[64] Carly K Schissel, Somesh Mohapatra, Justin M Wolfe, Colin M Fadzen, Kamela Bellovoda, Chia-Ling Wu, Jenna A Wood, Annika B Malmberg, Andrei Loas, Rafael Gómez-Bombarelli, et al. Deep learning to design nuclear-targeting abiotic miniproteins. *Nature Chemistry*, pages 1–9, 2021.

[65] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[66] Bo Qiao, Somesh Mohapatra, Jeffrey Lopez, Graham M Leverick, Ryoichi Tatara, Yoshiki Shibuya, Yivan Jiang, Arthur France-Lanord, Jeffrey C Grossman, Rafael Gómez-Bombarelli, et al. Quantitative mapping of molecular substituents to macroscopic properties enables predictive design of oligoethylene glycol-based lithium electrolytes. *ACS central science*, 6(7):1115–1128, 2020.

[67] Peter C St. John, Caleb Phillips, Travis W Kemper, A Nolan Wilson, Yanfei Guan, Michael F Crowley, Mark R Nimlos, and Ross E Larsen. Message-passing neural networks for high-throughput polymer screening. *The Journal of chemical physics*, 150(23):234111, 2019.

[68] Chee-Kong Lee, Chengqiang Lu, Yue Yu, Qiming Sun, Chang-Yu Hsieh, Shengyu Zhang, Qi Liu, and Liang Shi. Transfer learning with graph neural networks for optoelectronic properties of conjugated oligomers. *The Journal of Chemical Physics*, 154(2):024906, 2021.

[69] Lihua Chen, Ghanshyam Pilania, Rohit Batra, Tran Doan Huan, Chiho Kim, Christopher Kuenneth, and Rampi Ramprasad. Polymer informatics: Current status and critical next steps. *Materials Science and Engineering: R: Reports*, 144:100595, 2021.

[70] Greg Landrum et al. Rdkit: Open-source cheminformatics, 2006.

[71] Paola Gramatica. Whim descriptors of shape. *QSAR & Combinatorial Science*, 25(4):327–332, 2006.

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

[73] Scott Clark and Patrick Hayes. Sigopt web page. *SigOpt Web Page*, 2019.

[74] Tim Head, Gilles Louppe MechCoder, Iaroslav Shcherbatyi, et al. scikit-optimize/scikit-optimize: v0. 5.2. *Zenodo*, 2018.

[75] Panagiotis-Christos Kotsias, Josep Arús-Pous, Hongming Chen, Ola Engkvist, Christian Tyrchan, and Esben Jannik Bjerrum. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature Machine Intelligence*, 2(5):254–265, 2020.

[76] Somesh Mohapatra, Tzuhsiung Yang, and Rafael Gómez-Bombarelli. Reusability report: Designing organic photoelectronic molecules with descriptor conditional recurrent neural networks. *Nature Machine Intelligence*, 2(12):749–752, 2020.

[77] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[78] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[79] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[80] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[81] Somesh Mohapatra, Joyce An, and Rafael Gómez-Bombarelli. Chemistry-informed macromolecule graph representation for similarity computation and supervised learning. *arXiv preprint arXiv:2103.02565*, 2021.

[82] Gary C McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100, 2009.

[83] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013.

[84] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[85] Robert Dorfman. A formula for the gini coefficient. *The review of economics and statistics*, pages 146–149, 1979.

[86] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[87] Rudolph L Juliano. The delivery of therapeutic oligonucleotides. *Nucleic acids research*, 44(14):6518–6548, 2016.

[88] Tatiana A Slastnikova, Alexey V Ulasov, Andrey A Rosenkranz, and Alexander S Sobolev. Targeted intracellular delivery of antibodies: The state of the art. *Frontiers in pharmacology*, 9:1208, 2018.

[89] Ailing Fu, Rui Tang, Joseph Hardie, Michelle E Farkas, and Vincent M Rotello. Promises and pitfalls of intracellular delivery of proteins. *Bioconjugate chemistry*, 25(9):1602–1608, 2014.

[90] Shane Miersch and Sachdev S Sidhu. Intracellular targeting with engineered proteins. *F1000Research*, 5, 2016.

[91] Françoise Illien, Nicolas Rodriguez, Mehdi Amoura, Alain Joliot, Manjula Pallerla, Sophie Cribier, Fabienne Burlina, and Sandrine Sagan. Quantitative fluorescence spectroscopy and flow cytometry analyses of cell-penetrating peptides internalization pathways: optimization, pitfalls, comparison with mass spectrometry quantification. *Scientific reports*, 6(1):1–13, 2016.

[92] Prisca Boisguérin, Sébastien Deshayes, Michael J Gait, Liz O'Donovan, Caroline Godfrey, Corinne A Betts, Matthew JA Wood, and Bernard Lebleu. Delivery of therapeutic oligonucleotides with cell penetrating peptides. *Advanced drug delivery reviews*, 87:52–67, 2015.

[93] Corinne Betts, Amer F Saleh, Andrey A Arzumanov, Suzan M Hammond, Caroline Godfrey, Thibault Coursindel, Michael J Gait, and Matthew JA Wood. Pip6-pmo, a new generation of peptide-oligonucleotide conjugates with improved cardiac exon skipping activity for dmd treatment. *Molecular Therapy-Nucleic Acids*, 1:e38, 2012.

[94] Justin M Wolfe, Colin M Fadzen, Rebecca L Holden, Monica Yao, Gunnar J Hanson, and Bradley L Pentelute. Perfluoroaryl bicyclic cell-penetrating peptides for delivery of antisense oligonucleotides. *Angewandte Chemie*, 130(17): 4846–4849, 2018.

[95] Sebastian Spänig and Dominik Heider. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData mining*, 12(1):1–29, 2019.

[96] Ge Liu, Haoyang Zeng, Jonas Mueller, Brandon Carter, Ziheng Wang, Jonas Schilz, Geraldine Horny, Michael E Birnbaum, Stefan Ewert, and David K Gifford. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 36(7):2126–2133, 2020.

[97] Justin M Wolfe, Colin M Fadzen, Zi-Ning Choo, Rebecca L Holden, Monica Yao, Gunnar J Hanson, and Bradley L Pentelute. Machine learning to predict cell-penetrating peptides for antisense delivery. *ACS central science*, 4(4):512–520, 2018.

[98] Balachandran Manavalan, Sathiyamoorthy Subramaniyam, Tae Hwan Shin, Myeong Ok Kim, and Gwang Lee. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *Journal of proteome research*, 17(8):2715–2726, 2018.

[99] Ran Su, Jie Hu, Quan Zou, Balachandran Manavalan, and Leyi Wei. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Briefings in bioinformatics*, 21(2):408–420, 2020.

[100] William S Sanders, C Ian Johnston, Susan M Bridges, Shane C Burgess, and Kenneth O Willeford. Prediction of cell penetrating peptides by support vector machines. *PLoS computational biology*, 7(7):e1002101, 2011.

[101] Justin Mahoney Wolfe. *Peptide conjugation to enhance oligonucleotide delivery*. PhD thesis, Massachusetts Institute of Technology, 2018.

[102] Piyush Agrawal, Sherry Bhalla, Salman Sadullah Usmani, Sandeep Singh, Kumardeep Chaudhary, Gajendra PS Raghava, and Ankur Gautam. Cppsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic acids research*, 44(D1):D1098–D1103, 2016.

[103] Ankur Gautam, Harinder Singh, Atul Tyagi, Kumardeep Chaudhary, Rahul Kumar, Pallavi Kapoor, and GPS Raghava. Cppsite: a curated database of cell penetrating peptides. *Database*, 2012, 2012.

[104] Joel Ruben Antony Moniz and David Krueger. Nested lstms. In *Asian Conference on Machine Learning*, pages 530–544. PMLR, 2017.

[105] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

[106] William W Cohen, Pradeep Ravikumar, Stephen E Fienberg, et al. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, volume 3, pages 73–78. Citeseer, 2003.

[107] Enrique Audain, Yassel Ramos, Henning Hermjakob, Darren R Flower, and Yasset Perez-Riverol. Accurate estimation of isoelectric point of protein and peptide based on amino acid sequences. *Bioinformatics*, 32(6):821–827, 2016.

[108] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[109] Sebastien Deshayes, MC Morris, G Divita, and F Heitz. Cell-penetrating peptides: tools for intracellular delivery of therapeutics. *Cellular and Molecular Life Sciences CMLS*, 62(16):1839–1849, 2005.

[110] Eva M López-Vidal, Carly K Schissel, Somesh Mohapatra, Kamela Bellovoda, Chia-Ling Wu, Jenna A Wood, Annika B Malmberg, Andrei Loas, Rafael Gómez-Bombarelli, and Bradley L Pentelute. Deep learning enables discovery of a short nuclear targeting peptide for efficient delivery of antisense oligomers. *JACS Au*, 1(11):2009–2020, 2021.

[111] Kirk E Hevener, Russell Pesavento, JinHong Ren, Hyun Lee, Kiira Ratia, and Michael E Johnson. Hit-to-lead: Hit validation and assessment. In *Methods in Enzymology*, volume 610, pages 265–309. Elsevier, 2018.

[112] John J Bowling, William R Shadrick, Elizabeth C Griffith, and Richard E Lee. Going small: using biophysical screening to implement fragment based drug discovery. *Special Topics in Drug Discovery*, 25, 2016.

[113] H Gerhard Vogel, Wolfgang H Vogel, H Gerhard Vogel, Günter Müller, Jürgen Sandow, and Bernward A Schölkens. *Drug discovery and evaluation: pharmacological assays*, volume 2. Springer, 1997.

[114] Jonathan B Baell and Georgina A Holloway. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry*, 53(7):2719–2740, 2010.

[115] David J Payne, Michael N Gwynn, David J Holmes, and David L Pompliano. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature reviews Drug discovery*, 6(1):29–40, 2007.

[116] Simon Ng, Yu-Chi Juang, Arun Chandramohan, Hung Yi Kristal Kaan, Ahmad Sadruddin, Tsz Ying Yuen, Fernando J Ferrer-Gago, Xue'Er Cheryl Lee, Xi Liew, Charles W Johannes, et al. De-risking drug discovery of intracellular targeting peptides: screening strategies to eliminate false-positive hits. *ACS medicinal chemistry letters*, 11(10):1993–2001, 2020.

[117] 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.

[118] Natia Tsomaia. Peptide therapeutics: targeting the undruggable space. *European journal of medicinal chemistry*, 94:459–470, 2015.

[119] Alexander A Vinogradov, Yizhen Yin, and Hiroaki Suga. Macrocyclic peptides as drug candidates: recent progress and remaining challenges. *Journal of the American Chemical Society*, 141(10):4167–4181, 2019.

[120] Keld Fosgerau and Torsten Hoffmann. Peptide therapeutics: current status and future directions. *Drug discovery today*, 20(1):122–128, 2015.

[121] Alexander J Mijalis, Dale A Thomas, Mark D Simon, Andrea Adamo, Ryan Beaumont, Klavs F Jensen, and Bradley L Pentelute. A fully automated flow-based approach for accelerated peptide synthesis. *Nature chemical biology*, 13 (5):464–466, 2017.

[122] Zachary P Gates, Alexander A Vinogradov, Anthony J Quartararo, Anupam Bandyopadhyay, Zi-Ning Choo, Ethan D Evans, Kathryn H Halloran, Alexander J Mijalis, Surin K Mong, Mark D Simon, et al. Xenoprotein engineering via synthetic libraries. *Proceedings of the National Academy of Sciences*, 115(23): E5298–E5306, 2018.

[123] Jesper Lau, Paw Bloch, Lauge Schaffer, Ingrid Pettersson, Jane Spetzler, Jacob Kofoed, Kjeld Madsen, Lotte Bjerre Knudsen, James McGuire, Dorte Bjerre Steensgaard, et al. Discovery of the once-weekly glucagon-like peptide-1 (glp-1) analogue semaglutide. *Journal of medicinal chemistry*, 58(18):7370–7380, 2015.

[124] Hoi Yee Chow, Yue Zhang, Eilidh Matheson, and Xuechen Li. Ligation technologies for the synthesis of cyclic peptides. *Chemical reviews*, 119(17):9971–10001, 2019.

[125] Lauren RH Krumpe and Toshiyuki Mori. Potential of phage-displayed peptide library technology to identify functional targeting peptides. *Expert opinion on drug discovery*, 2(4):525–537, 2007.

[126] Tim Clackson and James A Wells. In vitro selection from protein and peptide libraries. *Trends in biotechnology*, 12(5):173–184, 1994.

[127] Xu-Dong Kong, Jun Moriya, Vanessa Carle, Florence Pojer, Luciano A Abriata, Kaycie Deyle, and Christian Heinis. De novo development of proteolytically resistant therapeutic peptides for oral administration. *Nature Biomedical Engineering*, 4(5):560–571, 2020.

[128] Richard W Roberts and Jack W Szostak. Rna-peptide fusions for the in vitro selection of peptides and proteins. *Proceedings of the National Academy of Sciences*, 94(23):12297–12302, 1997.

[129] Kristopher Josephson, Alonso Ricardo, and Jack W Szostak. mrna display: from basic principles to macrocycle drug discovery. *Drug Discovery Today*, 19 (4):388–399, 2014.

[130] Kit S Lam, Sydney E Salmon, Evan M Hersh, Victor J Hruby, Wieslaw M Kazmierski, and Richard J Knapp. A new type of synthetic peptide library for identifying ligand-binding activity. *Nature*, 354(6348):82–84, 1991.

[131] Ronald N Zuckermann, Janice M Kerr, Michael A Siani, Steven C Banville, and Daniel V Santi. Identification of highest-affinity ligands by affinity selection from equimolar peptide mixtures generated by robotic synthesis. *Proceedings of the National Academy of Sciences*, 89(10):4505–4509, 1992.

[132] Anthony J Quartararo, Zachary P Gates, Bente A Somsen, Nina Hartrampf, Xiyun Ye, Arisa Shimada, Yasuhiro Kajihara, Christian Ottmann, and Bradley L Pentelute. Ultra-large chemical libraries for the discovery of high-affinity peptide binders. *Nature communications*, 11(1):1–11, 2020.

[133] 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.

[134] Mari Yoshida, Trevor Hinkley, Soichiro Tsuda, Yousef M Abul-Haija, Roy T McBurney, Vladislav Kulikov, Jennifer S Mathieson, Sabrina Galiñanes Reyes, Maria D Castro, and Leroy Cronin. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem*, 4(3):533–543, 2018.

[135] Shaherin Basith, Balachandran Manavalan, Tae Hwan Shin, and Gwang Lee. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. *Medicinal research reviews*, 40(4):1276–1314, 2020.

[136] Sébastien Giguere, François Laviolette, Mario Marchand, Denise Tremblay, Sylvain Moineau, Xinxia Liang, Éric Biron, and Jacques Corbeil. Machine learning assisted design of highly active peptides for drug discovery. *PLoS computational biology*, 11(4):e1004074, 2015.

[137] Joseph M Cunningham, Grigoriy Koytiger, Peter K Sorger, and Mohammed AlQuraishi. Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nature methods*, 17(2):175–183, 2020.

[138] Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

[139] Xiaoshan M Shao, Rohit Bhattacharya, Justin Huang, IK Ashok Sivakumar, Collin Tokheim, Lily Zheng, Dylan Hirsch, Benjamin Kaminow, Ashton Omdahl, Maria Bonsack, et al. High-throughput prediction of mhc class i and ii neoantigens with mhcnuggets. *Cancer immunology research*, 8(3):396–408, 2020.

[140] Binbin Chen, Michael S Khodadoust, Niclas Olsson, Lisa E Wagar, Ethan Fast, Chih Long Liu, Yagmur Muftuoglu, Brian J Sworder, Maximilian Diehn, Ronald Levy, et al. Predicting hla class ii antigen presentation through integrated deep learning. *Nature biotechnology*, 37(11):1332–1343, 2019.

[141] Zhonghao Liu, Yuxin Cui, Zheng Xiong, Alierza Nasiri, Ansi Zhang, and Jianjun Hu. Deepseqpan, a novel deep convolutional neural network model for pan-specific class i hla-peptide binding affinity prediction. *Scientific reports*, 9(1): 1–10, 2019.

[142] Alexander A Vinogradov, Zachary P Gates, Chi Zhang, Anthony J Quartararo, Kathryn H Halloran, and Bradley L Pentelute. Library design-facilitated high-throughput sequencing of synthetic peptide libraries. *ACS combinatorial science*, 19(11):694–701, 2017.

[143] Michael AA Cox and Trevor F Cox. Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer, 2008.

[144] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

[145] Yu Hoshino, Haejoo Lee, and Yoshiko Miura. Interaction between synthetic particles and biomacromolecules: fundamental study of nonspecific interaction and design of nanoparticles that recognize target molecules. *Polymer Journal*, 46(9):537–545, 2014.

[146] Alexander L Satz, Remo Hochstrasser, and Ann C Petersen. Analysis of current dna encoded library screening data indicates higher false negative rates for numerically larger libraries. *ACS Combinatorial Science*, 19(4):234–238, 2017.

[147] Bjørn-Ivar Haukanes and Catrine Kvam. Application of magnetic beads in bioassays. *Bio/technology*, 11(1):60–63, 1993.

[148] Sergey Pevtsov, Irina Fedulova, Hamid Mirzaei, Charles Buck, and Xiang Zhang. Performance evaluation of existing de novo sequencing algorithms. *Journal of proteome research*, 5(11):3018–3028, 2006.

[149] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 17(20):2337–2342, 2003.

[150] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.

[151] Pavel Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855 (1-23):40, 2008.

[152] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[153] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):1–31, 2013.

[154] Selina Chu, Eamonn Keogh, David Hart, and Michael Pazzani. Iterative deepening dynamic time warping for time series. In *Proceedings of the 2002 SIAM International Conference on Data Mining*, pages 195–212. SIAM, 2002.

[155] Young-Seon Jeong, Myong K Jeong, and Olufemi A Omitaomu. Weighted dynamic time warping for time series classification. *Pattern recognition*, 44 (9):2231–2240, 2011.

[156] Cory Myers, Lawrence Rabiner, and Aaron Rosenberg. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(6):623–635, 1980.

[157] Toni Giorgino. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31:1–24, 2009.

[158] Duarte Folgado, Marília Barandas, Ricardo Matias, Rodrigo Martins, Miguel Carvalho, and Hugo Gamboa. Time alignment measurement for time series. *Pattern Recognition*, 81:268–279, 2018.

[159] Wayne C Koff, Dennis R Burton, Philip R Johnson, Bruce D Walker, Charles R King, Gary J Nabel, Rafi Ahmed, Maharaj K Bhan, and Stanley A Plotkin. Accelerating next-generation vaccine development for global disease prevention. *Science*, 340(6136):1232910, 2013.

[160] Keerthi Shetty and Patrick A Ott. Personal neoantigen vaccines for the treatment of cancer. *Annual Review of Cancer Biology*, 5:259–276, 2021.

[161] Ugur Sahin and Özlem Türeci. Personalized vaccines for cancer immunotherapy. *Science*, 359(6382):1355–1360, 2018.

[162] Zhuting Hu, Patrick A Ott, and Catherine J Wu. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nature Reviews Immunology*, 18(3): 168–182, 2018.

[163] Gerald P Linette and Beatriz M Carreno. Neoantigen vaccines pass the immunogenicity test. *Trends in molecular medicine*, 23(10):869–871, 2017.

[164] Ton N Schumacher and Robert D Schreiber. Neoantigens in cancer immunotherapy. *Science*, 348(6230):69–74, 2015.

[165] Janice S Blum, Pamela A Wearsch, and Peter Cresswell. Pathways of antigen processing. *Annual review of immunology*, 31:443–473, 2013.

[166] Jacques Neefjes, Marlieke LM Jongsma, Petra Paul, and Oddmund Bakke. Towards a systems understanding of mhc class i and mhc class ii antigen presentation. *Nature reviews immunology*, 11(12):823–836, 2011.

[167] Nilabh Shastri, Sylvain Cardinaud, Susan R Schwab, Thomas Serwold, and Jun Kunisawa. All the peptides that fit: the beginning, the middle, and the end of the mhc class i antigen-processing pathway. *Immunological reviews*, 207(1): 31–41, 2005.

[168] Kenneth L Rock, Ian A York, and Alfred L Goldberg. Post-proteasomal antigen processing for major histocompatibility complex class i presentation. *Nature immunology*, 5(7):670–677, 2004.

[169] Derin B Keskin, Annabelle J Anandappa, Jing Sun, Itay Tirosh, Nathan D Mathewson, Shuqiang Li, Giacomo Oliveira, Anita Giobbie-Hurder, Kristen Felt, Evisa Gjini, et al. Neoantigen vaccine generates intratumoral t cell responses in phase ib glioblastoma trial. *Nature*, 565(7738):234–239, 2019.

[170] Patrick A Ott, Zhuting Hu, Derin B Keskin, Sachet A Shukla, Jing Sun, David J Bozym, Wandi Zhang, Adrienne Luoma, Anita Giobbie-Hurder, Lauren Peter, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 547(7662):217–221, 2017.

[171] Frank Gambino Jr, Wanbo Tai, Denis Voronin, Yi Zhang, Xiujuan Zhang, Juan Shi, Xinyi Wang, Ning Wang, Lanying Du, and Liang Qiao. A vaccine inducing solely cytotoxic t lymphocytes fully prevents zika virus infection and fetal damage. *Cell Reports*, 35(6):109107, 2021.

[172] Fernando Rodriguez, Ling Ling An, Stephanie Harkins, Jie Zhang, Masayuki Yokoyama, Georg Widera, James T Fuller, Carrie Kincaid, Iain L Campbell, and J Lindsay Whitton. Dna immunization with minigenes: low frequency of memory cytotoxic t lymphocytes and inefficient antiviral protection are rectified by ubiquitination. *Journal of Virology*, 72(6):5174–5181, 1998.

[173] Siranush Sarkizova, Susan Klaeger, Phuong M Le, Letitia W Li, Giacomo Oliveira, Hasmik Keshishian, Christina R Hartigan, Wandi Zhang, David A Braun, Keith L Ligon, et al. A large peptidome dataset improves hla class i epitope prediction across most of the human population. *Nature biotechnology*, 38(2):199–209, 2020.

[174] Jennifer G Abelin, Derin B Keskin, Siranush Sarkizova, Christina R Hartigan, Wandi Zhang, John Sidney, Jonathan Stevens, William Lane, Guang Lan Zhang, Thomas M Eisenhaure, et al. Mass spectrometry profiling of hla-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity*, 46(2):315–326, 2017.

[175] Thomas Stranzl, Mette Voldby Larsen, Claus Lundegaard, and Morten Nielsen. Netctlpan: pan-specific mhc class i pathway epitope predictions. *Immunogenetics*, 62(6):357–368, 2010.

[176] Hadar Ella, Yuval Reiss, and Tommer Ravid. The hunt for degrons of the 26s proteasome. *Biomolecules*, 9(6):230, 2019.

[177] Andreas Bachmair, Daniel Finley, and Alexander Varshavsky. In vivo half-life of a protein is a function of its amino-terminal residue. *science*, 234(4773): 179–186, 1986.

[178] Hsueh-Chi Sherry Yen, Qikai Xu, Danny M Chou, Zhenming Zhao, and Stephen J Elledge. Global protein stability profiling in mammalian cells. *Science*, 322(5903):918–923, 2008.

[179] Hsueh-Chi Sherry Yen and Stephen J Elledge. Identification of scf ubiquitin ligase substrates by global protein stability profiling. *Science*, 322(5903):923–929, 2008.

[180] Amy E Rabideau and Bradley L Pentelute. A d-amino acid at the n-terminus of a protein abrogates its degradation by the n-end rule pathway. *ACS central science*, 1(8):423–430, 2015.

[181] Domniţa-Valeria Rusnac, Hsiu-Chuan Lin, Daniele Canzani, Karena X Tien, Thomas R Hinds, Ashley F Tsue, Matthew F Bush, Hsueh-Chi S Yen, and Ning Zheng. Recognition of the diglycine c-end degron by crl2klhdc2 ubiquitin ligase. *Molecular cell*, 72(5):813–822, 2018.

[182] Hsiu-Chuan Lin, Chi-Wei Yeh, Yen-Fu Chen, Ting-Ting Lee, Pei-Yun Hsieh, Domnita V Rusnac, Sung-Ya Lin, Stephen J Elledge, Ning Zheng, and Hsueh-Chi S Yen. C-terminal end-directed protein elimination by crl2 ubiquitin ligases. *Molecular cell*, 70(4):602–613, 2018.

[183] Itay Koren, Richard T Timms, Tomasz Kula, Qikai Xu, Mamie Z Li, and Stephen J Elledge. The eukaryotic proteome is shaped by e3 ubiquitin ligases targeting c-terminal degrons. *Cell*, 173(7):1622–1635, 2018.

[184] Frédéric Lévy, Lena Burri, Sandra Morel, Anne-Lise Peitrequin, Nicole Lévy, Angela Bachi, Ulf Hellman, Benoît J Van den Eynde, and Catherine Servis. The final n-terminal trimming of a subaminoterminal proline-containing hla class i-restricted antigenic peptide in the cytosol is mediated by two peptidases. *The Journal of Immunology*, 169(8):4161–4171, 2002.

[185] Abie Craiu, Tatos Akopian, Alfred Goldberg, and Kenneth L Rock. Two distinct proteolytic processes in the generation of a major histocompatibility complex class i-presented peptide. *Proceedings of the National Academy of Sciences*, 94 (20):10850–10855, 1997.

[186] Richard T Timms, Zhiqian Zhang, David Y Rhee, J Wade Harper, Itay Koren, and Stephen J Elledge. A glycine-specific n-degron pathway mediates the quality control of protein n-myristoylation. *Science*, 365(6448):eaaw4912, 2019.

[187] Jimmy D Ballard, R John Collier, and Michael N Starnbach. Anthrax toxin-mediated delivery of a cytotoxic t-cell epitope in vivo. *Proceedings of the National Academy of Sciences*, 93(22):12531–12534, 1996.

[188] Jingjing J Ling, Rocco L Policarpo, Amy E Rabideau, Xiaoli Liao, and Bradley L Pentelute. Protein thioester synthesis enabled by sortase. *Journal of the American Chemical Society*, 134(26):10749–10752, 2012.

[189] Anne-Catherine Bédard, Andrea Adamo, Kosi C Aroh, M Grace Russell, Aaron A Bedermann, Jeremy Torosian, Brian Yue, Klavs F Jensen, and Timothy F Jamison. Reconfigurable system for automated optimization of diverse chemical reactions. *Science*, 361(6408):1220–1225, 2018.

[190] BG Buchanan and D Waltz. Automating science. *Science*, 324(5923):43–44, 2009.

[191] Melanie Trobe and Martin D Burke. The molecular industrial revolution: automated synthesis of small molecules. *Angewandte Chemie International Edition*, 57(16):4192–4214, 2018.

[192] Junqi Li, Steven G Ballmer, Eric P Gillis, Seiko Fujii, Michael J Schmidt, Andrea ME Palazzolo, Jonathan W Lehmann, Greg F Morehouse, and Martin D Burke. Synthesis of many different types of organic small molecules using one automated process. *Science*, 347(6227):1221–1226, 2015.

[193] Eric M Woerly, Jahnabi Roy, and Martin D Burke. Synthesis of most polyene natural product motifs using just 12 building blocks and one coupling reaction. *Nature chemistry*, 6(6):484–491, 2014.

[194] Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen. Prediction of organic reaction outcomes using machine learning. *ACS central science*, 3(5):434–443, 2017.

[195] Loïc M Roch, Florian Häse, Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars PE Yunker, Jason E Hein, and Alán Aspuru-Guzik. Chemos: orchestrating autonomous experimentation. *Science Robotics*, 3(19):eaat5559, 2018.

[196] Sebastian Steiner, Jakob Wolf, Stefan Glatzel, Anna Andreou, Jarosław M Granda, Graham Keenan, Trevor Hinkley, Gerardo Aragon-Camarasa, Philip J Kitson, Davide Angelone, et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science*, 363(6423):eaav2211, 2019.

[197] Florian Hase, Loïc M Roch, Christoph Kreisbeck, and Alán Aspuru-Guzik. Phoenics: a bayesian optimizer for chemistry. *ACS central science*, 4(9):1134–1145, 2018.

[198] Benjamin P MacLeod, Fraser GL Parlane, Thomas D Morrissey, Florian Häse, Loïc M Roch, Kevan E Dettelbach, Raphaell Moreira, Lars PE Yunker, Michael B Rooney, Joseph R Deeth, et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20):eaaz8867, 2020.

[199] Connor W Coley, William H Green, and Klavs F Jensen. Machine learning in computer-aided synthesis planning. *Accounts of chemical research*, 51(5): 1281–1289, 2018.

[200] Zhenpeng Zhou, Xiaocheng Li, and Richard N Zare. Optimizing chemical reactions with deep reinforcement learning. *ACS central science*, 3(12):1337–1344, 2017.

[201] Hanyu Gao, Thomas J Struble, Connor W Coley, Yuran Wang, William H Green, and Klavs F Jensen. Using machine learning to predict suitable conditions for organic reactions. *ACS central science*, 4(11):1465–1476, 2018.

[202] Connor W Coley, Dale A Thomas III, Justin AM Lummiss, Jonathan N Jaworski, Christopher P Breen, Victor Schultz, Travis Hart, Joshua S Fishman, Luke Rogers, Hanyu Gao, et al. A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science*, 365(6453):eaax1566, 2019.

[203] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 2016.

[204] Nina Hartrampf, Azin Saebi, Mackenzie Poskus, Zachary P Gates, Alexander J Callahan, Amanda E Cowfer, Stephanie Hanna, Sarah Antilla, Carly K Schissel, Anthony J Quartararo, et al. Synthesis of proteins by automated flow chemistry. *Science*, 368(6494):980–987, 2020.

[205] Eric Atherton, Vivienne Woolley, and Robert C Sheppard. Internal association in solid phase peptide synthesis. synthesis of cytochrome c residues 66–104 on polyamide supports. *Journal of the Chemical Society, Chemical Communications*, 20:970–971, 1980.

[206] SBH Kent. Difficult sequences in stepwise peptide synthesis: common molecular origins in solution and solid phase? *ChemInform*, 18(14):no–no, 1987.

[207] Virender K Sarin, Stephen BH Kent, and RB Merrifield. Properties of swollen polymer networks. solvation and swelling of peptide-containing resins in solid-phase peptide synthesis. *Journal of the American Chemical Society*, 102(17): 5463–5470, 1980.

[208] RC de L Milton, Saskia CF Milton, and Paul A Adams. Prediction of difficult sequences in solid-phase peptide synthesis. *Journal of the American Chemical Society*, 112(16):6039–6046, 1990.

[209] WJ Van Woerkom and JW Van Nispen. Difficult couplings in stepwise solid phase peptide synthesis: predictable or just a guess? *International Journal of Peptide and Protein Research*, 38(2):103–113, 1991.

[210] JOY BEDFORD, CAROLYN HYDE, T Johnson, WEN JUN, D Owen, M Quibell, and RC Sheppard. Amino acid structure and "difficult sequences" in solid phase peptide synthesis. *International journal of peptide and protein research*, 40(3-4):300–307, 1992.

[211] Mark David Simon. *Fast flow biopolymer synthesis*. PhD thesis, Massachusetts Institute of Technology, 2017.

[212] Thomas J Lukas, Michael B Prystowsky, and Bruce W Erickson. Solid-phase peptide synthesis under continuous-flow conditions. *Proceedings of the National Academy of Sciences*, 78(5):2791–2795, 1981.

[213] Eric Atherton, Evelyn Brown, Robert C Sheppard, and Alan Rosevear. A physically supported gel polymer for low pressure, continuous flow solid phase reactions. application to solid phase peptide synthesis. *Journal of the Chemical Society, Chemical Communications*, 21:1151–1152, 1981.

[214] Linda R Cameron, Jill L Holder, Morten Meldal, and Robert C Sheppard. Peptide synthesis. part 13. feedback control in solid phase synthesis. use of fluorenylmethoxycarbonyl amino acid 3, 4-dihydro-4-oxo-1, 2, 3-benzotriazin-3-yl esters in a fully automated system. *Journal of the Chemical Society, Perkin Transactions 1*, 10:2895–2901, 1988.

[215] Alison Dryland and Robert C Sheppard. Peptide synthesis. part 8. a system for solid-phase synthesis under low pressure continuous flow conditions. *Journal of the Chemical Society, Perkin Transactions 1*, pages 125–137, 1986.

[216] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[217] Yahiya Y Syed. Eteplirsen: first global approval. *Drugs*, 76(17):1699–1704, 2016.

[218] Young-A Heo. Golodirsen: first approval. *Drugs*, 80(3):329–333, 2020.

[219] Matt Shirley. Casimersen: first approval. *Drugs*, 81(7):875–879, 2021.

[220] Sohita Dhillon. Viltolarsen: first approval. *Drugs*, 80(10):1027–1031, 2020.

[221] V Prakash. Spinraza—a rare disease success story. *Gene Therapy*, 24(9):497–497, 2017.

[222] Soheila Montazersaheb, Mohammad Saeid Hejazi, and Hojjatollah Nozad Charoudeh. Potential of peptide nucleic acids in future therapeutic applications. *Advanced pharmaceutical bulletin*, 8(4):551, 2018.

[223] Chiranjeev Sharma and Satish Kumar Awasthi. Versatility of peptide nucleic acids (pna s): Role in chemical biology, drug discovery, and origins of life. *Chemical Biology & Drug Design*, 89(1):16–37, 2017.

[224] Vadim V Demidov, Vladimir N Potaman, MD Frank-Kamenetskil, Michael Egholm, Ole Buchard, Søren H Sönnichsen, and Peter E Nlelsen. Stability of peptide nucleic acids in human serum and cellular extracts. *Biochemical pharmacology*, 48(6):1310–1313, 1994.

[225] Chengxi Li, Alexander J Callahan, Kruttika-Suhas Phadke, Bryan Bellaire, Charlotte E Farquhar, Genwei Zhang, Carly K Schissel, Alexander J Mijalis, Nina Hartrampf, Andrei Loas, et al. Automated fast-flow synthesis of peptide nucleic acid conjugates. *ChemRxiv*, 2021.

[226] Chengxi Li, Alex J Callahan, Mark D Simon, Kyle A Totaro, Alexander J Mijalis, Kruttika-Suhas Phadke, Genwei Zhang, Nina Hartrampf, Carly K Schissel, Ming Zhou, et al. Fully automated fast-flow synthesis of antisense phosphorodiamidate morpholino oligomers. *Nature Communications*, 12(1):1–8, 2021.

[227] Richard T Wang, Florian Barthelemy, Ann S Martin, Emilie D Douine, Ascia Eskin, Ann Lucas, Jenifer Lavigne, Holly Peay, Negar Khanlou, Lee Sweeney, et al. Dmd genotype correlations from the duchenne registry: Endogenous exon skipping is a factor in prolonged ambulation for individuals with a defined mutation subtype. *Human mutation*, 39(9):1193–1202, 2018.

[228] Ajit Varki. Biological roles of glycans. *Glycobiology*, 27(1):3–49, 2017.

[229] Matthew P Campbell, Louise Royle, Catherine M Radcliffe, Raymond A Dwek, and Pauline M Rudd. Glycobase and autogu: tools for hplc-based glycan analysis. *Bioinformatics*, 24(9):1214–1216, 2008.

[230] René Ranzinger, Stephan Herget, Claus-Wilhelm von der Lieth, and Martin Frank. Glycomedb—a unified database for carbohydrate structures. *Nucleic acids research*, 39(suppl_1):D373–D376, 2010.

[231] Pascal Gagneux, Markus Aebi, and Ajit Varki. Evolution of glycan diversity. *Essentials of Glycobiology [Internet]. 3rd edition*, 2017.

[232] Carolyn R Bertozzi and David Rabuka. *Structural basis of glycan diversity*. Cold Spring Harbor Laboratory Press, 2010.

[233] Daniel F Wyss and Gerhard Wagner. The structural role of sugars in glycoproteins. *Current opinion in biotechnology*, 7(4):409–416, 1996.

[234] Ari Helenius and Markus Aebi. Intracellular functions of n-linked glycans. *Science*, 291(5512):2364–2369, 2001.

[235] Rahul Raman, S Raguram, Ganesh Venkataraman, James C Paulson, and Ram Sasisekharan. Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nature methods*, 2(11):817–824, 2005.

[236] Nathaniel L Miller, Thomas Clark, Rahul Raman, and Ram Sasisekharan. Glycans in virus-host interactions: A structural perspective. *Frontiers in Molecular Biosciences*, 8:505, 2021.

[237] Ron Amon, Eliran Moshe Reuven, Shani Leviatan Ben-Arye, and Vered Padler-Karavani. Glycans in immune recognition and response. *Carbohydrate research*, 389:115–122, 2014.

[238] Saori Inafuku, Kousuke Noda, Maho Amano, Tetsu Ohashi, Chikako Yoshizawa, Wataru Saito, Miyuki Murata, Atsuhiro Kanda, Shin-Ichiro Nishimura, and Susumu Ishida. Alteration of n-glycan profiles in diabetic retinopathy. *Investigative ophthalmology & visual science*, 56(9):5316–5322, 2015.

[239] Sergei A Svarovsky and Lokesh Joshi. Cancer glycan biomarkers and their detection–past, present and future. *Analytical Methods*, 6(12):3918–3936, 2014.

[240] Stefan Gaunitz, Lars O Tjernberg, and Sophia Schedin-Weiss. The n-glycan profile in cortex and hippocampus is altered in alzheimer disease. *Journal of neurochemistry*, 159(2):292–304, 2021.

[241] Byeong Gwan Cho, Lucas Veillon, and Yehia Mechref. N-glycan profile of cerebrospinal fluids from alzheimer's disease patients using liquid chromatography with mass spectrometry. *Journal of proteome research*, 18(10):3770–3779, 2019.

[242] Gerald W Hart and Ronald J Copeland. Glycomics hits the big time. *Cell*, 143 (5):672–676, 2010.

[243] Richard D Cummings and J Michael Pierce. The challenge and promise of glycomics. *Chemistry & biology*, 21(1):1–15, 2014.

[244] Hyun Joo An, Scott R Kronewitter, Maria Lorna A de Leoz, and Carlito B Lebrilla. Glycomics and disease markers. *Current opinion in chemical biology*, 13(5-6):601–607, 2009.

[245] Jeremy E Turnbull and Robert A Field. Emerging glycomics technologies. *Nature Chemical Biology*, 3(2):74–77, 2007.

[246] Nicholas J Agard and Carolyn R Bertozzi. Chemical approaches to perturb, profile, and perceive glycans. *Accounts of chemical research*, 42(6):788–797, 2009.

[247] Saddam M Muthana, Christopher T Campbell, and Jeffrey C Gildersleeve. Modifications of glycans: biological significance and therapeutic opportunities. *ACS chemical biology*, 7(1):31–43, 2012.

[248] Ola Blixt, Steve Head, Tony Mondala, Christopher Scanlan, Margaret E Huflejt, Richard Alvarez, Marian C Bryan, Fabio Fazio, Daniel Calarese, James Stevens, et al. Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proceedings of the National Academy of Sciences*, 101 (49):17033–17038, 2004.

[249] Rahul Raman, Maha Venkataraman, Subu Ramakrishnan, Wei Lang, S Raguram, and Ram Sasisekharan. Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology*, 16(5):82R–90R, 2006.

[250] Daniel Bojar, Rani K Powers, Diogo M Camacho, and James J Collins. Sweetorigins: Extracting evolutionary information from glycans. *bioRxiv*, 2020.

[251] Daiji Fukagawa, Takeyuki Tamura, Atsuhiro Takasu, Etsuji Tomita, and Tatsuya Akutsu. A clique-based method for the edit distance between unordered trees and its application to analysis of glycan structures. *BMC bioinformatics*, 12(1):1–9, 2011.

[252] Hui Sun Lee, Sunhwan Jo, Srayanta Mukherjee, Sang-Jun Park, Jeffrey Skolnick, Jooyoung Lee, and Wonpil Im. Gs-align for glycan structure alignment and similarity measurement. *Bioinformatics*, 31(16):2653–2659, 2015.

[253] Eric J Carpenter, Shaurya Seth, Noel Yue, Russell Greiner, and Ratmir Derda. Glynet: A multi-task neural network for predicting protein-glycan interactions. *bioRxiv*, 2021.

[254] Lachlan Coff, Jeffrey Chan, Paul A Ramsland, and Andrew J Guy. Identifying glycan motifs using a novel subtree mining approach. *BMC bioinformatics*, 21 (1):1–18, 2020.

[255] Sharath R Cholleti, Sanjay Agravat, Tim Morris, Joel H Saltz, Xuezheng Song, Richard D Cummings, and David F Smith. Automated motif discovery from glycan array data. *Omics: a journal of integrative biology*, 16(10):497–512, 2012.

[256] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[257] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[258] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

[259] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick F. Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, 2016.

[260] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. *arXiv:1704.01212v2*, 2017.

[261] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020.

[262] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2019.

[263] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph Attention Networks. *arXiv:1710.10903*, 2017.

[264] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv:1909.01315*, 2019.

[265] Rebekka Burkholz, John Quackenbush, and Daniel Bojar. Using graph convolutional neural networks to learn a representation for glycans. *Cell Reports*, 35 (11):109251, 2021.

[266] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv:1703.01365*, 2017.

[267] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *arXiv:1704.02685*, 2017.

[268] Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Y Wang, Wei Qian, Kevin Mccloskey, Lucy Colwell, and Alexander Wiltschko. Evaluating Attribution for Graph Neural Networks. *Proceeedings of the Neural Information Processing Systems Conference*, 2020.

[269] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[270] Jon Lundstrøm, Emma Korhonen, Frédérique Lisacek, and Daniel Bojar. Lectinoracle: A generalizable deep learning model for lectin–glycan binding prediction. *Advanced Science*, 9(1):2103807, 2022.

[271] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[272] Eric V Anslyn, Dennis A Dougherty, et al. *Modern physical organic chemistry*. University science books, 2006.

[273] Louis P Hammett. The effect of structure upon the reactions of organic compounds. benzene derivatives. *Journal of the American Chemical Society*, 59(1): 96–103, 1937.

[274] Tu Le, V Chandana Epa, Frank R Burden, and David A Winkler. Quantitative structure–property relationship modeling of diverse materials properties. *Chemical reviews*, 112(5):2889–2919, 2012.

[275] Jolene P Reid and Matthew S Sigman. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nature Reviews Chemistry*, 2(10):290–305, 2018.

[276] Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin A Blood-Forsythe, Hyun Sik Chae, Markus Einzinger, Dong-Gwang Ha, Tony Wu, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials*, 15(10):1120–1127, 2016.

[277] Mingjiang Zhong, Rui Wang, Ken Kawamoto, Bradley D Olsen, and Jeremiah A Johnson. Quantifying the impact of molecular defects on polymer network elasticity. *Science*, 353(6305):1264–1268, 2016.

[278] Mohammad Vatankhah-Varnosfaderani, William FM Daniel, Matthew H Everhart, Ashish A Pandya, Heyi Liang, Krzysztof Matyjaszewski, Andrey V Dobrynin, and Sergei S Sheiko. Mimicking biological stress–strain behaviour with synthetic elastomers. *Nature*, 549(7673):497–501, 2017.

[279] Daniel T Hallinan Jr and Nitash P Balsara. Polymer electrolytes. *Annual review of materials research*, 43:503–525, 2013.

[280] Shuting Feng, Mingjun Huang, Jessica R Lamb, Wenxu Zhang, Ryoichi Tatara, Yirui Zhang, Yun Guang Zhu, Collin F Perkinson, Jeremiah A Johnson, and Yang Shao-Horn. Molecular design of stable sulfamide-and sulfonamide-based electrolytes for aprotic li-o2 batteries. *Chem*, 5(10):2630–2641, 2019.

[281] Mingjun Huang, Shuting Feng, Wenxu Zhang, Jeffrey Lopez, Bo Qiao, Ryoichi Tatara, Livia Giordano, Yang Shao-Horn, and Jeremiah A Johnson. Design of s-substituted fluorinated aryl sulfonamide-tagged (s-fast) anions to enable new solvate ionic liquids for battery applications. *Chemistry of Materials*, 31(18): 7558–7564, 2019.

[282] Mingjun Huang, Shuting Feng, Wenxu Zhang, Livia Giordano, Mao Chen, Chibueze V Amanchukwu, Robinson Anandakathir, Yang Shao-Horn, and Jeremiah A Johnson. Fluorinated aryl sulfonimide tagged (fast) salts: modular synthesis and structure–property relationships for battery applications. *Energy & Environmental Science*, 11(5):1326–1334, 2018.

[283] Ardeshir Baktash, James C Reid, Qinghong Yuan, Tanglaw Roman, and Debra J Searles. Shaping the future of solid-state electrolytes through computational modeling. *Advanced Materials*, 32(18):1908041, 2020.

[284] Dominic Bresser, Sandrine Lyonnard, Cristina Iojoiu, Lionel Picard, and Stefano Passerini. Decoupling segmental relaxation and ionic conductivity for lithium-ion polymer electrolytes. *Molecular Systems Design & Engineering*, 4(4):779–792, 2019.

[285] Yuki Kato, Satoshi Hori, Toshiya Saito, Kota Suzuki, Masaaki Hirayama, Akio Mitsui, Masao Yonemura, Hideki Iba, and Ryoji Kanno. High-power all-solid-state batteries using sulfide superionic conductors. *Nature Energy*, 1(4):1–7, 2016.

[286] Jeffrey Lopez, David G Mackanic, Yi Cui, and Zhenan Bao. Designing polymers for advanced battery chemistries. *Nature Reviews Materials*, 4(5):312–330, 2019.

[287] Kazuhide Ueno, Kazuki Yoshida, Mizuho Tsuchiya, Naoki Tachikawa, Kaoru Dokko, and Masayoshi Watanabe. Glyme–lithium salt equimolar molten mixtures: concentrated solutions or solvate ionic liquids? *The Journal of Physical Chemistry B*, 116(36):11323–11331, 2012.

[288] Richard Hooper, Leslie J Lyons, Marie K Mapes, Douglas Schumacher, David A Moline, and Robert West. Highly conductive siloxane polymers. *Macromolecules*, 34(4):931–936, 2001.

[289] X Andrieu, JF Fauvarque, A Goux, T Hamaide, R M'hamdi, and T Vicedo. Solid polymer electrolytes based on statistical poly (ethylene oxide-propylene oxide) copolymers. *Electrochimica acta*, 40(13-14):2295–2299, 1995.

[290] Katherine P Barteau, Martin Wolffs, Nathaniel A Lynd, Glenn H Fredrickson, Edward J Kramer, and Craig J Hawker. Allyl glycidyl ether-based polymer electrolytes for room temperature lithium batteries. *Macromolecules*, 46(22): 8988–8994, 2013.

[291] Renaud Bouchet, Sébastien Maria, Rachid Meziane, Abdelmaula Aboulaich, Livie Lienafa, Jean-Pierre Bonnet, Trang NT Phan, Denis Bertin, Didier Gigmes, Didier Devaux, et al. Single-ion bab triblock copolymers as highly efficient electrolytes for lithium-metal batteries. *Nature materials*, 12(5):452–457, 2013.

[292] Takeshi Niitani, Mikiya Shimada, Kiyoshi Kawamura, Kaoru Dokko, Young-Ho Rho, and Kiyoshi Kanamura. Synthesis of li+ ion conductive peo-pst block copolymer electrolyte with microphase separation structure. *Electrochemical and Solid-State Letters*, 8(8):A385, 2005.

[293] Yoshiki Shibuya, Ryoichi Tatara, Yivan Jiang, Yang Shao-Horn, and Jeremiah A Johnson. Brush-first romp of poly (ethylene oxide) macromonomers of varied length: impact of polymer architecture on thermal behavior and li+ conductivity. *Journal of Polymer Science Part A: Polymer Chemistry*, 57(3): 448–455, 2019.

[294] Masayoshi Watanabe, Takuro Hirakimoto, Shunsuke Mutoh, and Atsushi Nishimoto. Polymer electrolytes derived from dendritic polyether macromonomers. *Solid State Ionics*, 148(3-4):399–404, 2002.

[295] JT Dudley, DP Wilkinson, G Thomas, R LeVae, S Woo, H Blom, C Horvath, MW Juzkow, B Denis, P Juric, et al. Conductivity of electrolytes for rechargeable lithium batteries. *Journal of power sources*, 35(1):59–82, 1991.

[296] Patrik Johansson, Mark A Ratner, and Duward F Shriver. The influence of inert oxide fillers on poly (ethylene oxide) and amorphous poly (ethylene oxide) based polymer electrolytes. *The Journal of Physical Chemistry B*, 105(37): 9016–9021, 2001.

[297] R Tanaka, M Sakurai, H Sekiguchi, H Mori, T Murayama, and T Ooyama. Lithium ion conductivity in polyoxyethylene/polyethylenimine blends. *Electrochimica acta*, 46(10-11):1709–1715, 2001.

[298] Mark A Ratner and Duward F Shriver. Ion transport in solvent-free polymers. *Chemical reviews*, 88(1):109–124, 1988.

[299] Michael A Webb, Yukyung Jung, Danielle M Pesko, Brett M Savoie, Umi Yamamoto, Geoffrey W Coates, Nitash P Balsara, Zhen-Gang Wang, and Thomas F Miller III. Systematic computational and experimental investigation of lithium-ion transport mechanisms in polyester-based polymer electrolytes. *ACS central science*, 1(4):198–205, 2015.

[300] Enrique D Gomez, Ashoutosh Panday, Edward H Feng, Vincent Chen, Gregory M Stone, Andrew M Minor, Christian Kisielowski, Kenneth H Downing, Oleg Borodin, Grant D Smith, et al. Effect of ion distribution on conductivity of block copolymer electrolytes. *Nano letters*, 9(3):1212–1216, 2009.

[301] Vaidyanathan Sethuraman, Santosh Mogurampelly, and Venkat Ganesan. Multiscale simulations of lamellar ps–peo block copolymers doped with lipf6 ions. *Macromolecules*, 50(11):4542–4554, 2017.

[302] Danielle M Pesko, Michael A Webb, Yukyung Jung, Qi Zheng, Thomas F Miller III, Geoffrey W Coates, and Nitash P Balsara. Universal relationship

between conductivity and solvation-site connectivity in ether-based polymer electrolytes. *Macromolecules*, 49(14):5244–5255, 2016.

[303] Yuki Kato, Kentaro Suwa, Hiromasa Ikuta, Yoshiharu Uchimoto, Masataka Wakihara, Shoichi Yokoyama, Takeshi Yabe, and Masahiro Yamamoto. Influence of lewis acidic borate ester groups on lithium ionic conduction in polymer electrolytes. *Journal of Materials Chemistry*, 13(2):280–285, 2003.

[304] Brett M Savoie, Michael A Webb, and Thomas F Miller III. Enhancing cation diffusion and suppressing anion diffusion via lewis-acidic polymer electrolytes. *The journal of physical chemistry letters*, 8(3):641–646, 2017.

[305] Qi Zheng, Danielle M Pesko, Brett M Savoie, Ksenia Timachova, Alexandra L Hasan, Mackensie C Smith, Thomas F Miller III, Geoffrey W Coates, and Nitash P Balsara. Optimizing ion transport in polyether-based electrolytes for lithium batteries. *Macromolecules*, 51(8):2847–2858, 2018.

[306] John Newman, Karen E Thomas, Hooman Hafezi, and Dean R Wheeler. Modeling of lithium-ion batteries. *Journal of power sources*, 119:838–843, 2003.

[307] Alejandro A Franco. Multiscale modelling and numerical simulation of rechargeable lithium ion batteries: concepts, methods and challenges. *Rsc Advances*, 3 (32):13027–13058, 2013.

[308] Mitchell T Ong, Osvalds Verners, Erik W Draeger, Adri CT Van Duin, Vincenzo Lordi, and John E Pask. Lithium ion solvation and diffusion in bulk organic electrolytes from first-principles and classical reactive molecular dynamics. *The Journal of Physical Chemistry B*, 119(4):1535–1545, 2015.

[309] Mitchell T Ong, Harsh Bhatia, Attila G Gyulassy, Erik W Draeger, Valerio Pascucci, Peer-Timo Bremer, Vincenzo Lordi, and John E Pask. Complex ion dynamics in carbonate lithium-ion battery electrolytes. *The Journal of Physical Chemistry C*, 121(12):6589–6595, 2017.

[310] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, 8(4):3192–3203, 2017.

[311] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O Anatole Von Lilienfeld, Klaus-Robert Muller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters*, 6(12):2326–2331, 2015.

[312] Harikrishna Sahu, Weining Rao, Alessandro Troisi, and Haibo Ma. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Advanced Energy Materials*, 8(24):1801032, 2018.

[313] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

[314] Kaid C Harper and Matthew S Sigman. Three-dimensional correlation of steric and electronic free energy relationships guides asymmetric propargylation. *Science*, 333(6051):1875–1878, 2011.

[315] Sophia G Robinson, Yichao Yan, Koen H Hendriks, Melanie S Sanford, and Matthew S Sigman. Developing a predictive solubility model for monomeric and oligomeric cyclopropenium-based flow battery catholytes. *Journal of the American Chemical Society*, 141(26):10171–10176, 2019.

[316] Jens Kehlet Nørskov, Thomas Bligaard, Jan Rossmeisl, and Claus Hviid Christensen. Towards the computational design of solid catalysts. *Nature chemistry*, 1(1):37–46, 2009.

[317] Morihiro Saito, Shinya Yamada, Taro Ishikawa, Hiromi Otsuka, Kimihiko Ito, and Yoshimi Kubo. Factors influencing fast ion transport in glyme-based electrolytes for rechargeable lithium–air batteries. *RSC advances*, 7(77):49031–49040, 2017.

[318] Nana Arai, Hikari Watanabe, Tsuyoshi Yamaguchi, Shiro Seki, Kazuhide Ueno, Kaoru Dokko, Masayoshi Watanabe, Yasuo Kameda, Richard Buchner, and Yasuhiro Umebayashi. Dynamic chelate effect on the li+-ion conduction in solvate ionic liquids. *The Journal of Physical Chemistry C*, 123(50):30228–30233, 2019.

[319] Ce Zhang, Kazuhide Ueno, Azusa Yamazaki, Kazuki Yoshida, Heejoon Moon, Toshihiko Mandai, Yasuhiro Umebayashi, Kaoru Dokko, and Masayoshi Watanabe. Chelate effects in glyme/lithium bis (trifluoromethanesulfonyl) amide solvate ionic liquids. i. stability of solvate cations and correlation with electrolyte properties. *The Journal of Physical Chemistry B*, 118(19):5144–5153, 2014.

[320] Gabriela Horwitz, Matías Factorovich, Javier Rodriguez, Daniel Laria, and Horacio R Corti. Ionic transport and speciation of lithium salts in glymes: experimental and theoretical results for electrolytes of interest for lithium–air batteries. *ACS omega*, 3(9):11205–11215, 2018.

[321] Xiang Chen, Yun-Ke Bai, Chen-Zi Zhao, Xin Shen, and Qiang Zhang. Lithium bonds in lithium batteries. *Angewandte Chemie*, 132(28):11288–11291, 2020.

[322] Kyle M Diederichsen, Hilda G Buss, and Bryan D McCloskey. The compensation effect in the vogel–tammann–fulcher (vtf) equation for polymer-based electrolytes. *Macromolecules*, 50(10):3831–3840, 2017.

[323] Lei Liu and Qing-Xiang Guo. Isokinetic relationship, isoequilibrium relationship, and enthalpy- entropy compensation. *Chemical Reviews*, 101(3):673–696, 2001.

[324] James R MacCallum and Colin Angus Vincent. *Polymer electrolyte reviews*, volume 2. Springer Science & Business Media, 1989.

[325] Chibueze V Amanchukwu, Zhiao Yu, Xian Kong, Jian Qin, Yi Cui, and Zhenan Bao. A new class of ionically conducting fluorinated ether electrolytes with high electrochemical stability. *Journal of the American Chemical Society*, 142(16): 7393–7403, 2020.

[326] Roland Yawo Getor, Nishikant Mishra, and Amar Ramudhin. The role of technological innovation in plastic production within a circular economy framework. *Resources, Conservation and Recycling*, 163:105094, 2020.

[327] Richard C Thompson, Shanna H Swan, Charles J Moore, and Frederick S Vom Saal. Our plastic age, 2009.

[328] Edward Kosior and Jonathan Mitchell. Current industry position on plastic production and recycling. In *Plastic Waste and Recycling*, pages 133–162. Elsevier, 2020.

[329] Jennifer A Brandon, William Jones, and Mark D Ohman. Multidecadal increase in plastic particles in coastal ocean sediments. *Science advances*, 5(9):eaax0587, 2019.

[330] Emily J North and Rolf U Halden. Plastics and environmental health: the road ahead. *Reviews on environmental health*, 28(1):1–8, 2013.

[331] Ian D Robertson, Mostafa Yourdkhani, Polette J Centellas, Jia En Aw, Douglas G Ivanoff, Elyas Goli, Evan M Lloyd, Leon M Dean, Nancy R Sottos, Philippe H Geubelle, et al. Rapid energy-efficient manufacturing of polymers and composites via frontal polymerization. *Nature*, 557(7704):223–227, 2018.

[332] Peyton Shieh, Wenxu Zhang, Keith EL Husted, Samantha L Kristufek, Boya Xiong, David J Lundberg, Jet Lem, David Veysset, Yuchen Sun, Keith A Nelson, et al. Cleavable comonomers enable degradable, recyclable thermoset plastics. *Nature*, 583(7817):542–547, 2020.

[333] Hilmar Körner, Atsushi Shiota, and Christopher K Ober. Controlled-order thermosets for electronic packaging. In *Imaging and image analysis applications for plastics*, pages 283–287. Elsevier, 1999.

[334] Theo Pesenti and Julien Nicolas. 100th anniversary of macromolecular science viewpoint: Degradable polymers from radical ring-opening polymerization: Latest advances, new directions, and ongoing challenges. *ACS Macro Letters*, 9(12):1812–1835, 2020.

[335] Songqi Ma and Dean C Webster. Degradable thermosets based on labile bonds or linkages: A review. *Progress in Polymer Science*, 76:65–110, 2018.

[336] KEL Husted, P Shieh, DJ Lundberg, SL Kristufek, and JA Johnson. Molecularly designed additives for chemically deconstructable thermosets without compromised thermomechanical properties. *ACS Macro Letters*, 10(7):805–810, 2021.

[337] Shingo Otsuka, Isao Kuwajima, Junko Hosoya, Yibin Xu, and Masayoshi Yamazaki. Polyinfo: Polymer database for polymeric materials design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, pages 22–29. IEEE, 2011.

[338] Ruimin Ma and Tengfei Luo. Pi1m: a benchmark database for polymer informatics. *Journal of Chemical Information and Modeling*, 60(10):4684–4690, 2020.

[339] Yun Zhang and Xiaojie Xu. Machine learning glass transition temperature of polymers. *Heliyon*, 6(10):e05055, 2020.

[340] Lei Tao, Vikas Varshney, and Ying Li. Benchmarking machine learning models for polymer informatics: An example of glass transition temperature. *Journal of Chemical Information and Modeling*, 61(11):5395–5413, 2021.

[341] Lei Tao, Guang Chen, and Ying Li. Machine learning discovery of high-temperature polymers. *Patterns*, 2(4):100225, 2021.

[342] Stephen Wu, Yukiko Kondo, Masa-aki Kakimoto, Bin Yang, Hironao Yamada, Isao Kuwajima, Guillaume Lambard, Kenta Hongo, Yibin Xu, Junichiro Shiomi, et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *Npj Computational Materials*, 5(1):1–11, 2019.

[343] Lihua Chen, Chiho Kim, Rohit Batra, Jordan P Lightstone, Chao Wu, Zongze Li, Ajinkya A Deshmukh, Yifei Wang, Huan D Tran, Priya Vashishta, et al. Frequency-dependent dielectric constant prediction of polymers using machine learning. *npj Computational Materials*, 6(1):1–9, 2020.

[344] Shruti Venkatram, Rohit Batra, Lihua Chen, Chiho Kim, Madeline Shelton, and Rampi Ramprasad. Predicting crystallization tendency of polymers using multifidelity information fusion and machine learning. *The Journal of Physical Chemistry B*, 124(28):6046–6054, 2020.

[345] Rohit Batra, Hanjun Dai, Tran Doan Huan, Lihua Chen, Chiho Kim, Will R Gutekunst, Le Song, and Rampi Ramprasad. Polymers for extreme conditions designed using syntax-directed variational autoencoders. *Chemistry of Materials*, 32(24):10489–10500, 2020.

[346] Tyler R Long, Robert M Elder, Erich D Bain, Kevin A Masser, Timothy W Sirk, H Yu Jian, Daniel B Knorr, and Joseph L Lenhart. Influence of molecular weight between crosslinks on the mechanical properties of polymers formed via ring-opening metathesis. *Soft Matter*, 14(17):3344–3360, 2018.

[347] Sukdeb Saha, Yakov Ginzburg, Illya Rozenberg, Olga Iliashevsky, Amos Ben-Asuly, and N Gabriel Lemcoff. Cross-linked romp polymers based on odourless dicyclopentadiene derivatives. *Polymer Chemistry*, 7(18):3071–3075, 2016.

[348] Robert M Elder, Tyler R Long, Erich D Bain, Joseph L Lenhart, and Timothy W Sirk. Mechanics and nanovoid nucleation dynamics: effects of polar functionality in glassy polymer networks. *Soft Matter*, 14(44):8895–8911, 2018.

[349] Gregory S Constable, Alan J Lesser, and E Bryan Coughlin. Morphological and mechanical evaluation of hybrid organic- inorganic thermoset copolymers of dicyclopentadiene and mono-or tris (norbornenyl)-substituted polyhedral oligomeric silsesquioxanes. *Macromolecules*, 37(4):1276–1282, 2004.

[350] Brian J Rohde, Kim Mai Le, Ramanan Krishnamoorti, and Megan L Robertson. Thermoset blends of an epoxy resin and polydicyclopentadiene. *Macromolecules*, 49(23):8960–8970, 2016.

[351] Zhilong He, Jiacheng Sun, Li Zhang, Yaqi Wang, Huiyan Ren, and Jin-Biao Bao. Synergistic reinforcing and toughening of polydicyclopentadiene nanocomposites with low loadings vinyl-functionalized multi-walled carbon nanotubes. *Polymer*, 153:287–294, 2018.

[352] Yuval Vidavsky, Yotam Navon, Yakov Ginzburg, Moshe Gottlieb, and N Gabriel Lemcoff. Thermal properties of ruthenium alkylidene-polymerized dicyclopentadiene. *Beilstein Journal of Organic Chemistry*, 11(1):1469–1474, 2015.

[353] Daniel B Knorr Jr, Kevin A Masser, Robert M Elder, Timothy W Sirk, Mark D Hindenlang, H Yu Jian, Adam D Richardson, Steven E Boyd, William A Spurgeon, and Joseph L Lenhart. Overcoming the structural versus energy dissipation trade-off in highly crosslinked polymer networks: Ultrahigh strain rate response in polydicyclopentadiene. *Composites Science and Technology*, 114: 17–25, 2015.

[354] Anurag Jha, Anand Chandrasekaran, Chiho Kim, and Rampi Ramprasad. Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures. *Modelling and Simulation in Materials Science and Engineering*, 27(2):024002, 2019.

[355] Luis A Miccio and Gustavo A Schwartz. Localizing and quantifying the intramonomer contributions to the glass transition temperature using artificial neural networks. *Polymer*, 203:122786, 2020.

[356] Luis A Miccio and Gustavo A Schwartz. From chemical structure to quantitative polymer properties prediction through convolutional neural networks. *Polymer*, 193:122341, 2020.

[357] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge

Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[358] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131, 2018.

[359] Josep Arús-Pous, Thomas Blaschke, Silas Ulander, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Exploring the gdb-13 chemical space using deep generative models. *Journal of cheminformatics*, 11(1):1–14, 2019.

[360] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

[361] Johannes Hachmann, Roberto Olivares-Amaya, Adrian Jinich, Anthony L Appleton, Martin A Blood-Forsythe, László R Seress, Carolina Román-Salgado, Kai Trepte, Sule Atahan-Evrenk, Süleyman Er, et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry–the harvard clean energy project. *Energy & Environmental Science*, 7(2):698–704, 2014.

[362] Christian A Gueymard. The sun's total and spectral irradiance for solar energy applications and solar radiation models. *Solar energy*, 76(4):423–453, 2004.

[363] Steven A Lopez, Edward O Pyzer-Knapp, Gregor N Simm, Trevor Lutzow, Kewei Li, Laszlo R Seress, Johannes Hachmann, and Alán Aspuru-Guzik. The harvard organic photovoltaic dataset. *Scientific data*, 3(1):1–7, 2016.

[364] Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10 (12):3567–3572, 2019.

[365] 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.

[366] Sean Molesky, Zin Lin, Alexander Y Piggott, Weiliang Jin, Jelena Vucković, and Alejandro W Rodriguez. Inverse design in nanophotonics. *Nature Photonics*, 12(11):659–670, 2018.

[367] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361 (6400):360–365, 2018.

[368] Keith T Butler, Jarvist M Frost, Jonathan M Skelton, Katrine L Svane, and Aron Walsh. Computational materials design of crystalline solids. *Chemical Society Reviews*, 45(22):6138–6146, 2016.

[369] Alex Zunger. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry*, 2(4):1–16, 2018.

[370] Benjamin Sanchez-Lengeling, Carlos Outeiral, Gabriel L Guimaraes, and Alan Aspuru-Guzik. Optimizing distributions over molecular space. an objective-reinforced generative adversarial network for inverse-design chemistry (organic). *ChemRxiv*, 2017.

[371] Baekjun Kim, Sangwon Lee, and Jihan Kim. Inverse design of porous materials using artificial neural networks. *Science advances*, 6(1):eaax9324, 2020.

[372] Zhenpeng Yao, Benjamín Sánchez-Lengeling, N Scott Bobbitt, Benjamin J Bucior, Sai Govind Hari Kumar, Sean P Collins, Thomas Burns, Tom K Woo, Omar K Farha, Randall Q Snurr, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1):76–86, 2021.

[373] Juhwan Noh, Jaehoon Kim, Helge S Stein, Benjamin Sanchez-Lengeling, John M Gregoire, Alan Aspuru-Guzik, and Yousung Jung. Inverse design of solid-state materials via a continuous representation. *Matter*, 1(5):1370–1384, 2019.

[374] Juhwan Noh, Geun Ho Gu, Sungwon Kim, and Yousung Jung. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chemical Science*, 11(19):4871–4881, 2020.

[375] Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G Aberle, Shijing Sun, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 2021.

[376] Stefan Grimme. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *Journal of chemical theory and computation*, 15(5):2847–2862, 2019.

[377] C Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, et al. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv:2010.09435*, 2020.

[378] Alastair DG Lawson, Malcolm MacCoss, and Jag P Heer. Importance of rigidity in designing small molecule drugs to tackle protein–protein interactions (ppis) through stabilization of desired conformers: Miniperspective. *Journal of Medicinal Chemistry*, 61(10):4283–4289, 2017.

[379] Julian Tirado-Rives and William L Jorgensen. Contribution of conformer focusing to the uncertainty in predicting free energies for protein- ligand binding. *Journal of medicinal chemistry*, 49(20):5880–5884, 2006.

[380] Sereina Riniker and Gregory A Landrum. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling*, 55(12):2562–2574, 2015.

[381] Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22(14):7169–7192, 2020.

[382] Stefan Grimme, Fabian Bohle, Andreas Hansen, Philipp Pracht, Sebastian Spicher, and Marcel Stahn. Efficient quantum chemical calculation of structure ensembles and free energies for nonrigid molecules. *The Journal of Physical Chemistry A*, 125(19):4039–4054, 2021.

[383] Paul CD Hawkins and Anthony Nicholls. Conformer generation with omega: learning from the data set and the analysis of failures. *Journal of chemical information and modeling*, 52(11):2919–2936, 2012.

[384] Octavian-Eugen Ganea, Lagnajit Pattanaik, Connor W Coley, Regina Barzilay, Klavs F Jensen, William H Green, and Tommi S Jaakkola. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *arXiv preprint arXiv:2106.07802*, 2021.

[385] Gregor Simm and Jose Miguel Hernandez-Lobato. A generative model for molecular distance geometry. In *International Conference on Machine Learning*, pages 8949–8958. PMLR, 2020.

[386] Minkai Xu, Shitong Luo, Yoshua Bengio, Jian Peng, and Jian Tang. Learning neural generative dynamics for molecular conformation generation. In *International Conference on Learning Representations*, 2020.

[387] Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. *arXiv preprint arXiv:2105.03902*, 2021.

[388] Minkai Xu, Wujie Wang, Shitong Luo, Chence Shi, Yoshua Bengio, Rafael Gomez-Bombarelli, and Jian Tang. An end-to-end framework for molecular conformation generation via bilevel programming. *arXiv preprint arXiv:2105.07246*, 2021.

[389] Tristan Aumentado-Armstrong. Latent molecular optimization for targeted therapeutic design. *arXiv preprint arXiv:1809.02032*, 2018.

[390] Miha Skalic, José Jiménez, Davide Sabbadin, and Gianni De Fabritiis. Shape-based generative modeling for de novo drug design. *Journal of chemical information and modeling*, 59(3):1205–1214, 2019.

[391] Miha Skalic, Davide Sabbadin, Boris Sattarov, Simone Sciabola, and Gianni De Fabritiis. From target to drug: generative modeling for the multimodal structure-based ligand design. *Molecular pharmaceutics*, 16(10):4282–4291, 2019.

[392] Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Learning a continuous representation of 3d molecular structures with deep generative models. *arXiv preprint arXiv:2010.08687*, 2020.

[393] Tomohide Masuda, Matthew Ragoza, and David Ryan Koes. Generating 3d molecular structures conditional on a receptor binding site with deep generative models. *arXiv preprint arXiv:2010.14442*, 2020.

[394] Mingyuan Xu, Ting Ran, and Hongming Chen. De novo molecule design through the molecular generative model conditioned by 3d information of protein binding sites. *Journal of Chemical Information and Modeling*, 61(7):3240–3254, 2021.

[395] Fergus Imrie, Anthony R Bradley, Mihaela van der Schaar, and Charlotte M Deane. Deep generative models for 3d linker design. *Journal of chemical information and modeling*, 60(4):1983–1995, 2020.

[396] Fergus Imrie, Thomas E Hadfield, Anthony R Bradley, and Charlotte M Deane. Deep generative design with 3d pharmacophoric constraints. *Chemical science*, 12(43):14577–14589, 2021.

[397] Simon Axelrod and Rafael Gomez-Bombarelli. Geom: Energy-annotated molecular conformations for property prediction and molecular generation. *arXiv preprint arXiv:2006.05531*, 2020.

[398] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671, 2019.

[399] R Todeschini and P Gramatica. The whim theory: New 3d molecular descriptors for qsar in environmental modelling. *SAR and QSAR in Environmental Research*, 7(1-4):89–115, 1997.

[400] S Prasanna and RJ Doerksen. Topological polar surface area: a useful descriptor in 2d-qsar. *Current medicinal chemistry*, 16(1):21–41, 2009.

[401] Lemont B Kier and Lowell H Hall. An electrotopological-state index for atoms in molecules. *Pharmaceutical research*, 7(8):801–807, 1990.

[402] Miguel García-Ortegón, Gregor NC Simm, Austin J Tripp, José Miguel Hernández-Lobato, Andreas Bender, and Sergio Bacallado. Dockstring: easy molecular docking yields better benchmarks for ligand design. *arXiv preprint arXiv:2110.15486*, 2021.