npj computational materials

www.nature.com/npjcompumats

PERSPECTIVE OPEN

Check for updates

Sizing up feature descriptors for macromolecular machine learning with polymeric biomaterials

Samantha Stuart[1], Jeffrey Watchorn[2] and Frank X. Gu[1,2]✉

It has proved challenging to represent the behavior of polymeric macromolecules as machine learning features for biomaterial interaction prediction. There are several approaches to this representation, yet no consensus for a universal representational framework, in part due to the sensitivity of biomacromolecular interactions to polymer properties. To help navigate the process of feature engineering, we provide an overview of popular classes of data representations for polymeric biomaterial machine learning while discussing their merits and limitations. Generally, increasing the accessibility of polymeric biomaterial feature engineering knowledge will contribute to the goal of accelerating clinical translation from biomaterials discovery.

npj Computational Materials (2023)9:102 ; https://doi.org/10.1038/s41524-023-01040-5

INTRODUCTION

The selection of feature descriptors to encode a dataset for machine learning is one of the most important decisions underlying model quality, as different data representations can yield different interpretations of training data by the model[1,2]. With this, appropriate descriptors should be chosen with care and intention at the outset of a machine-learning project. Small molecules, as a function of their constrained sizes and structure, can be represented as standardized numeric descriptors for simulation, molecular property prediction, and virtual screening[3]. The ability to encode small molecules numerically in part provided an essential foundation for the chemoinformatics domain to achieve data-driven research success in small molecule drug discovery[4].

Inspired by small molecule success, machine learning frameworks for studying polymers often use feature descriptors based on the attributes of drug-like small molecules[4,5]. The intrinsic limitation of applying small-molecule-based feature representations to biomaterials is that small molecule descriptors lack the ability to accommodate the heterogeneity of polymer properties, which are drawn from combinations of polymer chemical, physical, and topological attributes[3,6]. Further, alterations in these macromolecular properties can yield significant changes in predictive target outcomes, such as a polymer's resulting interactions in biological media[7]. Similar examples including changes in polymer molecular weight, degree of polymerization, co-polymer, branching, chirality, nanostructure, synthesis technique, storage conditions, environmental conditions, polydispersity, and side chain regularity have all been shown to impact interaction outcomes[7–12]. In view of the limitations of small molecule descriptors for representing polymeric biomaterials, there is a clear need for dedicated macromolecular descriptors that facilitate the training of representative predictive models in this domain.

Unfortunately, it has proved challenging to generally represent the behavior of polymer biomaterials and their interactions with other biological macromolecules for machine learning. There are several popular approaches to polymer representation, including domain-specific descriptors, molecular fingerprints, string descriptors, and graph representations (Fig. 1), albeit there is no recognized consensus for an optimal representation across the biomacromolecular problem spaces[5,7,13–16]. Where one size does *not* fit all, researchers must independently consider the attributes of their dataset and objectives of their research project to contextually identify how descriptors can positively drive their resulting model towards robust predictive performance, a process known in machine learning as feature engineering.

Feature engineering encompasses refining and structuring raw input data into a relevant data structure that enables machine learning. Generally, this process is problem specific and relies heavily on domain (a priori) knowledge of a given problem. Domain knowledge guides the design and selection of features relevant to training accurate machine learning-based models. As an example, for polymer and biopolymer systems, variables such as molecular weight, degree of polymerization, and representation of the polymer sequence are often selected as features[17]. While it is possible to use these variables as a feature vector directly, the engineering process often involves some transformation to either improve machine readability or benefit predictive power. Examples include denoting presence or absence via one-hot encoding, binning a numerical value into categorical ranges, improving learning speed and preventing numerical overflow with scaling, or applying information compression algorithms such as principal component analysis or UMAP. Where deep learning algorithms are used, the input feature vector is again transformed through stacked layers of neurons, the ideal numbers of which vary in accordance with the dataset, and are optimized during the training process.

The exploratory process of feature engineering, particularly in research domains that have not been extensively studied with machine learning, can be time-consuming. Hence, reducing the effort involved in selecting descriptors of polymer properties and biomacromolecular interactions provides an important foundation for the generation of quick and unbiased research insights with machine learning. Increased dissemination of polymeric biomaterial feature engineering knowledge will serve to reduce time spent on the feature engineering project phase, and contribute to the shared goal of accelerating the path from materials discovery to biomaterial clinical translation and commercialization[18–21].

[1]Institute of Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada. [2]Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, Ontario, Canada. ✉email: f.gu@utoronto.ca

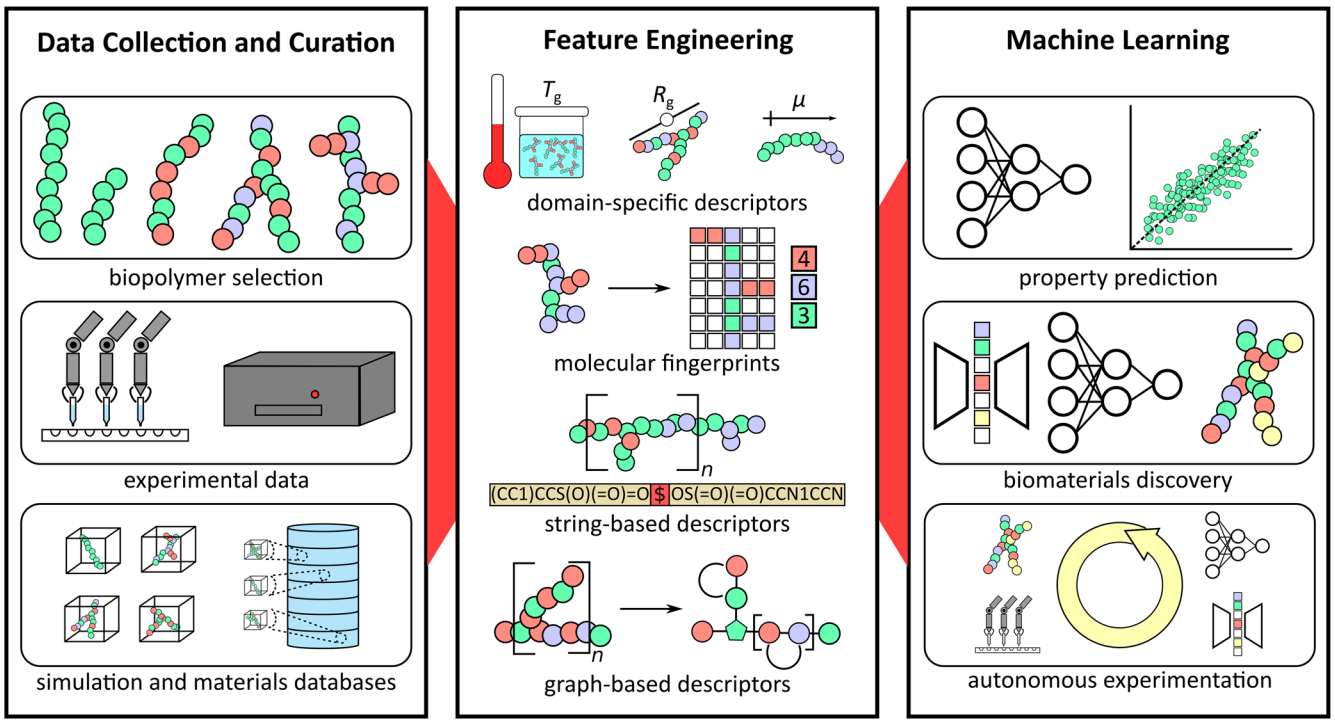

Fig. 1 A visual representation of the general process for applying machine learning to biomacromolecular modeling and discovery. Feature engineering functions as a central pillar between data collection and modeling, hence careful consideration of descriptor frameworks can have dramatic influence on model performance.

Towards this aim, in this perspective, we provide overviews, as well as discussion of the advantages and limitations, of different classes of macromolecular data representations applicable to polymeric biomaterial machine learning frameworks. Many polymeric biomaterial machine learning research efforts focus on interaction prediction tasks, such as modeling how polymers will interact with a target protein, or a biological environment containing other macromolecules. Modeling interaction outcomes ultimately informs the selection of polymers for use in medical devices, which by design induce such biological interactions. Additionally, biomaterials are often composites of multiple material types (polymers, proteins, nucleic acids, peptides, etc.). While it is convenient to express proteins, peptides, and nucleic acids using their primary sequence, this is not true for polymeric materials, hence representing such a composite is a fundamental challenge to biomaterials development[22].

With this in mind, we have focussed these discussions on the four most popular classes of macromolecular representation applicable to such polymer and biomaterials research: domain-specific descriptors, molecular fingerprints, string descriptors, and graph descriptors, described at a high level in Table 1.

Throughout this review, we highlight examples of research applying polymer data representations that can contribute to achieving predictive biomaterial design; such that polymers and biopolymers can be proactively selected for use in a biomaterial to achieve targeted biological outcomes. We hope that this perspective will benefit researchers seeking greater technical context on feature engineering for predictive polymer biomaterial design, as well as researchers in computer science seeking greater domain context on the challenges researchers face when building predictive models of large polymer systems for biomaterials engineering.

Domain-specific macromolecular descriptors

Research focused on training supervised learning models and interpreting their learned understandings through feature importance have clarified complex biological interaction mechanisms, and inspired research directions in the macromolecular biosciences[23–27]. Such works have been conducted from expert curated datasets on the order of 100 data points or greater[11,28–30] and apply problem-specific modeling features designed by researchers intimately familiar with the physics of the domain. Altogether, supervised learning followed by feature importance analysis of problem-specific macromolecular descriptors is an excellent use case for machine learning, where there are high-quality datasets describing multivariate problem spaces[25,31]. There are a growing number of examples of these works to draw inspiration from across research domains that employ macromolecular biomaterials. In particular, employing analytical characterization methods in conjunction with supervised learning has proved imperative for success in deconvoluting complex behaviors.

Analytical descriptors derived from mass spectroscopy are one powerful example in this regard. Proteomic descriptors from mass spectroscopy, when combined with supervised learning, have extracted wide-ranging biomechanistic insights, including the detection of Alzheimer's disease from nanoparticle protein coronas[32–34]. In one such example from nanomedicine, supervised learning and mass spectrometry were combined to accurately predict the biodistribution of nanomaterials in vivo using protein quantities present in the protein corona of PEGylated gold nanoparticles[23]. The analytical descriptors used as inputs for the neural network in this analysis were the label-free quantitative intensities from mass spectroscopy of proteins isolated from the surface of 8, 15, 35, 50, and 80 nm gold nanoparticles, over the course of 24 h of circulation in rats ($t = 1$, 2, 4, 8, and 24 h). Outputs were the resulting half-life, spleen gold accumulation, and liver gold accumulation of the nanoparticles as measured by inductively coupled plasma-mass spectrometry (ICP-MS). The workflow mapping the descriptors and outputs in this work is illustrated in Fig. 2. Other analytically derived nanoparticle design attributes can also be applied as descriptors, such as size, zeta

Table 1. Summary of popular macromolecular descriptor classes.

Descriptor Class	Overview	Implementation	Labor-Intensiveness
Domain-specific	Task-specific analytical measurements of the underlying system	Tabular encoding of experimental conditions, physical properties, physics-based simulations, analytical measurement results in accordance with domain knowledge	High (unless data collection is automated or variables known a priori)
Fingerprint	Vector encoding of the macromolecular chemistry in the training dataset	Tabular encoding of vectorized system, adoption of pre-existing frameworks where compatible	Low
String Representation	Encoded molecular structures using a predefined chemically complete knowledge framework	Incorporate framework in tabular data (ex. SMILES, BIGSMILES, SELFIES, etc.)	Medium
Graph Representation	Encoding and attribution of molecular systems using a graph data structure at a predefined level of abstraction	Custom encoding of nodes and edges for a graph learning task, adoption of pre-existing frameworks where compatible	Medium

potential, molecular weight, and associated experimental conditions such as cell type, exposure time, exposure route, and concentration[6]. Direct descriptors of the elemental composition of self-assembling monolayers (SAMs) have also been successful for interaction prediction tasks on these macromolecular assemblies, specifically %C, %H, %O, %N, total number of atoms, and number of O-H, C-C, C-O, C-N, and C=O bonds in the SAM[28].

Multiple modes of analytical descriptors may also be required to accurately model a biomacromolecular system. For example, augmenting experimentally derived data with high throughput physics-based simulation data can be considered, particularly in domains where molecular docking and molecular dynamics are applicable. These physics-based modeling techniques are helpful for establishing physical constraints for inverse design problems, even in cases for biological systems or complex materials design (such as in biomaterials design) where the physical models describing these processes are not well defined[35]. Moreover, these simulations can help to combat data sparsity, especially for physical parameters that would be difficult to determine experimentally, while providing an end-to-end quantification of overall model uncertainty[35]. Some examples of parameters of interest include the expected free energy of binding in protein-ligand interaction screening[36]; the diffusivity, probability of sequestration or vascular adhesion of nanocarriers for cancer drug delivery[37]; and dipole moment, polarizability, and hydrogen bond donor/acceptor ability for polymer solubility prediction tasks[38,39]. A final consideration for physics-based models for biopolymer prediction tasks is the length scale that the model should operate. For example, quantum chemical calculations to model electronic properties are accurate for small molecules, but often neglect considerations for polymeric materials such as conformation or morphology[40]. Coarse-grained modeling presents as an appropriate trade-off for polymer and biopolymer systems, where the goal is to represent higher-resolution systems with fewer degrees of freedom, which enables simulations at length and time scales more representative of biopolymer systems[41]. For example, coarse-grained simulations of biopolymeric galactomannans were able to accurately model the static structure, solution viscosity, and radius of gyration of guar gum gels[42].

In terms of other domain-specific feature generators, nuclear magnetic resonance (NMR) and time of flight secondary ion mass spectrometry can also be purposefully applied to generate analytical macromolecular descriptors of complex biomaterial interactions[43–45] which are well suited for supervised machine learning[36,46]. One such study focused on both polymer discovery and increasing mechanistic understanding of polymers for optimal ribonucleoprotein (RNP) delivery[11]. The authors experimentally screened a library of 43 copolymers to map nine polymer descriptors to their association with toxicity and gene editing efficiency, including: polyplex radius, polymer % cationic monomer (determined by NMR), molecular weight, pK_a, polymer hydrophobicity, RNP binding affinity, Hill coefficient, N/P ratio (i.e., nitrogen to phosphate group ratio), and charge density[11]. The trained random forest classifier identified polymer design attributes important for gene editing efficiency that the authors found counterintuitive, in particular flagging hydrophobically driven cooperative deprotonation as a promising mechanism for delivery[11]. Taken together, methodologies that parallelize polymer synthesis, high throughput screening, and multi-variate modeling are expected to continue driving results in biomaterial interaction prediction tasks[1,7,16,47].

Finally, designed analytical descriptors have also shown promise in inverse design tasks. A supervised learning model was accurately trained using a 117-sample dataset to predict the cloud point of poly(2-oxazoline), a polymer with emerging applications in biomaterials[48,49], using gradient-boosted decision trees and custom descriptors comprising varying ratios of four select monomer units, and molecular weight[30]. Molecular weight and composition ratio were identified as descriptors using domain knowledge in data curation. Interestingly, despite training from a relatively small dataset, the model in this work successfully executed polymer inverse design by synthesizing 17 de novo polymers with targeted cloud points between 37 and 80 °C with errors consistent with experimental ranges[30]. Inverse design tasks are often conducted using string descriptors, however, this among other similar works suggests that the targeted selection of physical property macromolecular descriptors can allow for the inverse design of macromolecule systems within narrow, well-defined, chemical spaces learned by supervised models[30,50–53].

Despite the success of experimentally derived biomacromolecular descriptors for supervised learning, there are some challenges with this form of data representation. For one, any syntax or semantics underlying the behavior of biomacromolecules is not preserved as an integral part of the descriptors. Alternative approaches directed towards including higher-order semantic molecular information for biomacromolecules are described further on in the language-based, and graph representation section.

Reducing the initial feature set into the final set of independent descriptors most relevant to the modeling task is also challenging. Directly encoded domain-specific feature vectors describing polymeric biomaterials can possess important variables which are highly intercorrelated (i.e., Pearson Correlation Coefficient > 0.85). Domain knowledge must be applied in such cases to decide whether to remove one of the intercorrelated variables through a standardized procedure such as the Least Absolute Shrinkage and Selection Operator (LASSO), or alternatively apply a dimensionality reduction technique such as principal component analysis (PCA) which enforces no linear intercorrelation between input variables. Several works make use of the LASSO method for

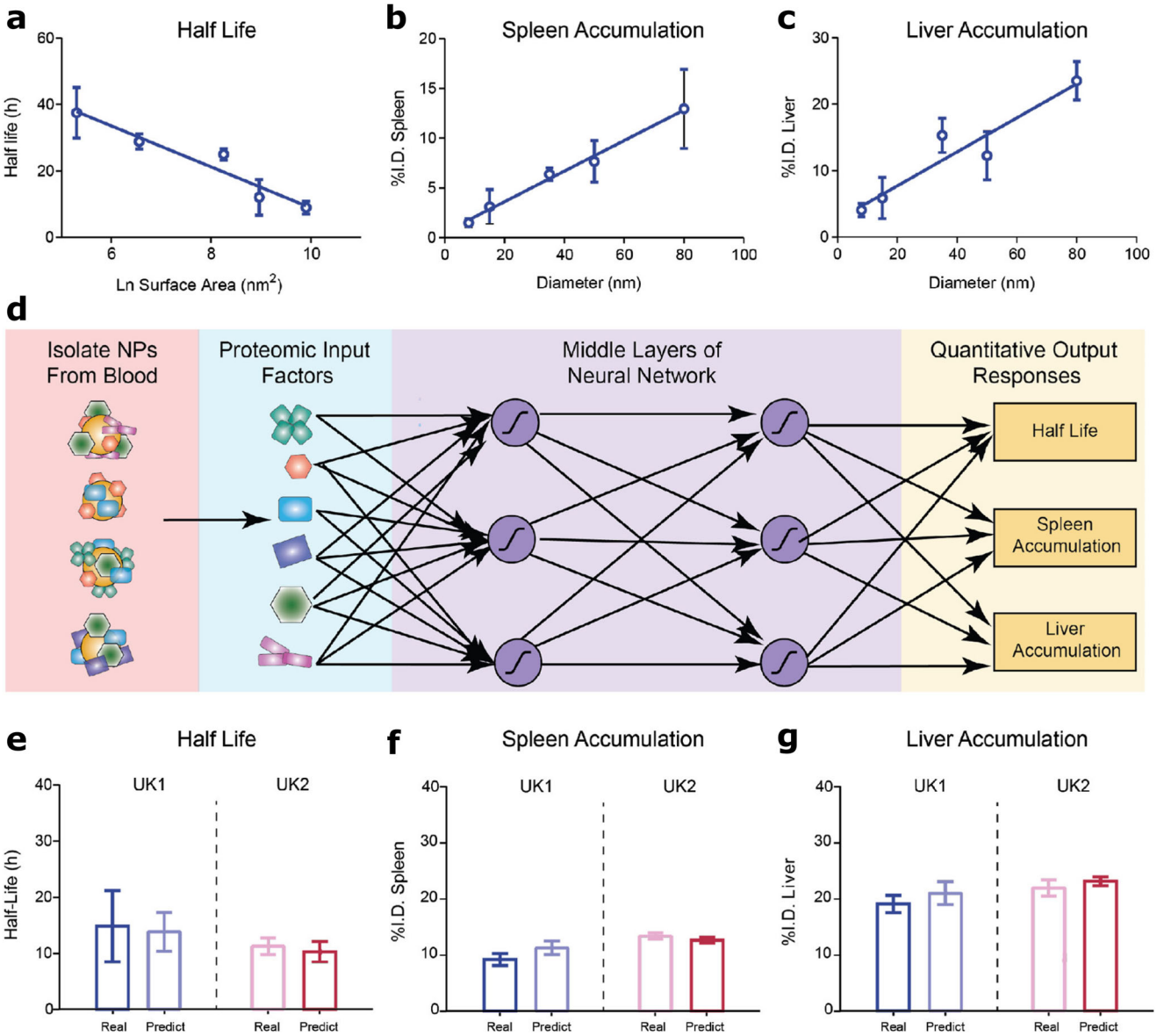

Fig. 2 Supervised learning descriptors derived from mass spectroscopy predict in vivo fate of PEGylated gold nanoparticles. The half-life (**a**), spleen accumulation (**b**), and liver accumulation (**c**) of five nanoparticle sizes were applied as target labels to train an artificial neural network to map proteomic input descriptors to in vivo nanoparticle fate (**d**). The model generalized successfully in predicting the properties of two unknown (UK) nanoparticles for their half-life (**e**), spleen accumulation (**f**), and liver accumulation (**g**). $n = 3$, error bars indicate standard deviation. Adapted with permission from ref. [23] (copyright American Chemical Society, 2019).

feature selection to eliminate the least relevant descriptors to the prediction task[1,23]. However, in biomaterial polymer informatics, the subject properties of the model can be intercorrelated from fundamentally networked attributes, rendering dropping such information before modeling undesirable[54]. In these cases, one may use PCA to remove feature intercorrelations while limiting information loss at the expense of some direct interpretability of the resulting feature importance[54]. Some of this interpretability can be retained by examining the factor loadings obtained through PCA as they correlate with the contribution of a given input variable to the model[55]. These loadings can also be used as an unbiased means of deriving insights from complex data over the baseline of manual interpretation[25]. Ultimately, problem context best informs the choice of methodology for biomaterial polymer feature selection, as it does with feature engineering on the whole.

In fact, evidence from benchmarking featurization strategies for polymer property prediction suggests not only that problem context dictates which feature engineering strategy will be the best performing, but that the predictive performance of a model can degrade after applying other feature engineering strategies to a fixed problem context[17]. Polymer size is one such example of a modeling feature for which problem context dictates its predictive significance.

In the benchmark, two datasets were contrasted for their performance sensitivity to polymer size as a modeling feature in a regression task. In regression, the resulting mean absolute error (MAE) benchmarks the predictive performance of the trained model on the task. In one dataset, including polymer size as a feature decreased MAE by 50% for each of its three property prediction tasks, a marked improvement. The other dataset, including polymer size did not yield any statistically significant reductions in MAE, indicating no effect. Domain knowledge

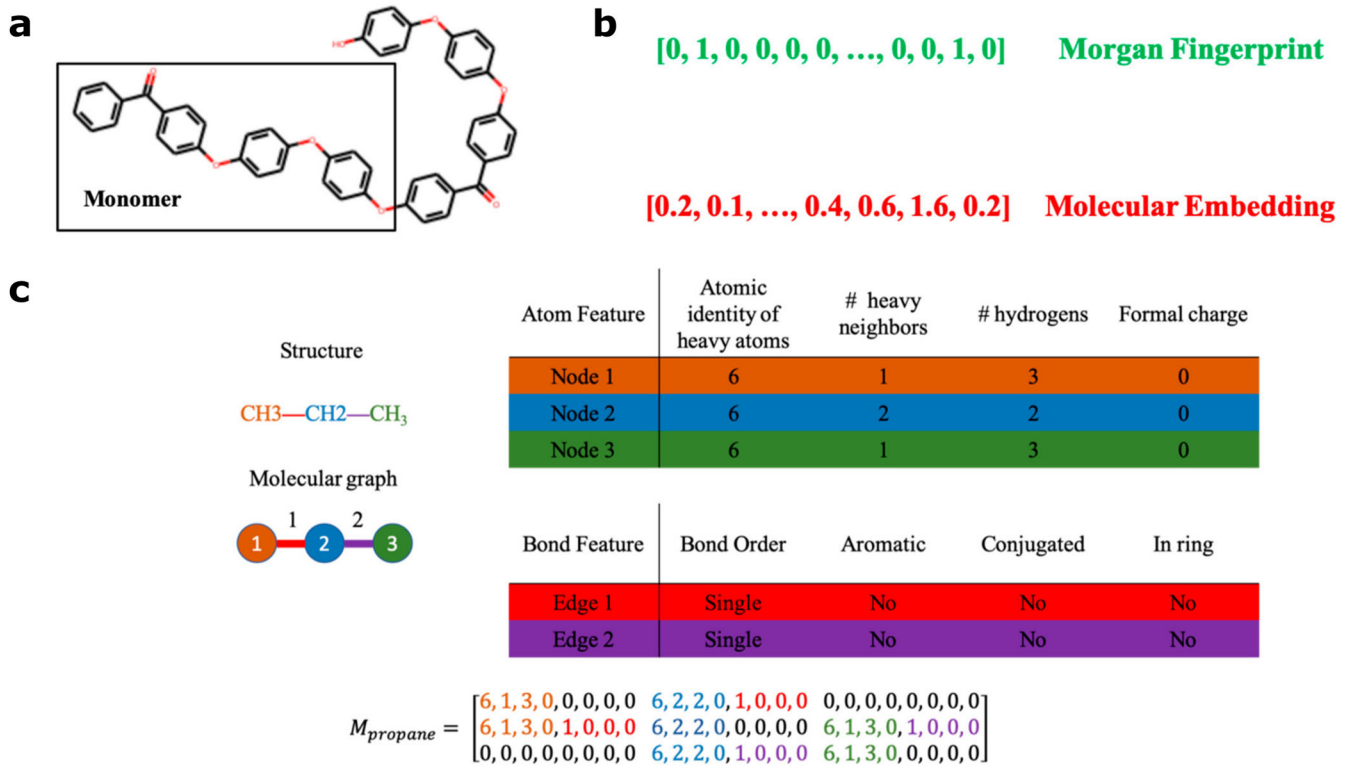

Fig. 3 Outline of molecular fingerprinting workflow applied to a two-unit polymer structure. a An exemplary two-monomer-unit polymer subset. **b** Binarized chemical attributes (Morgan Fingerprint) juxtaposed with an intra-molecular property enriched fingerprint (Molecular Embedding). **c** Explanatory diagrams mapping chemical structures to their characteristic atom features, bond features, and the resulting globally abstracted Molecular Graph. Adapted with permission from ref. [5] (copyright American Chemical Society, 2019).

suggests two problem context factors could underlie the discrepancy in the importance of the polymer size feature. First, the target variables of the first dataset were each sensitive to the polymer size ranges characterized in the training data (20–600 constitutional units), while the prediction target of the second dataset was not sensitive at the resolution being modeled (mean = 7770 g/mol, std = 1100 g/mol). Second, that measurement noise obfuscated any effect that was measured in the training data of the second dataset, which was collected by a different means than the first dataset. These contextual factors demonstrate the common bottlenecks encountered in biomaterial polymer dataset curation. Specifically, it is exceedingly challenging, and often laborious, in biomaterials design tasks to identify both polymer design space ranges that correspond to interaction behaviors of interest, and characterize those ranges reproducibly at scale, with measurement error that does not obfuscate the desired signal[17].

Formulating prediction tasks whereby challenges in data curation can be overcome remains an ongoing focus in the biomaterial polymer research domain. In addition to polymer size, there are innumerable design variables in polymer biomaterial development (physical, chemical, topological, etc.) whose curation will similarly impact the success of the featurization strategy in a given problem context. The challenges inherent to data curation in polymeric biomaterial design begets the premise of this work, that one size does not fit all in selecting a feature engineering strategy across predictive tasks. In a domain where problem context dictates the best feature engineering strategy, focusing on optimizing features to suit the variable-target mappings within a curated biopolymer dataset will trump the application of a generic strategy across problem domains that are not similarly curated, and where data points are scarce[56].

The scarcity of datapoints due to the labor-intensiveness of manual data collection imposes another limitation in feature engineering that merits noting, “the curse of dimensionality.” That being, the phenomenon in deep learning where as a feature vector dimension increases, the greater the number of datapoints in the dataset required to train a model. In biomaterial domains where interaction phenomena are typically unmapped, one may consider including every available parameter as a feature to increase the probability of mapping data to the prediction target. However, the curse of dimensionality enforces an upper limit in feature vector dimensionality as predicated by the amount of data able to be collected for the predictive task. Alternatively, one may augment a small, experimentally curated domain descriptor dataset using information from open-access databases or other research studies. However, pooling datapoints as such imbues mixed variance levels in the dataset from different experimental conditions, which again risks obscuring the desired objective function for the task[22]. All such factors in data curation similarly obfuscate the ability to draw standalone comparisons of model architectures based on different biomaterial datasets. Specifically, variance in performance can be attributed to any one or combination of data curation, feature engineering, or model design and training workflows. As such, across datasets, a one-size “fits all” approach has not yet accommodated all varieties of domain-specific factors inherent to polymeric biomaterials design.

In sum, domain descriptors can yield extremely informative feature mappings between experimental parameters and target variables for a wide variety of prediction tasks in polymeric biomaterial interaction prediction. However, the highly laborious nature of data collection and data curation remains an obstacle to scaling domain descriptors, and underscores the importance of both combinatorial polymer chemistry[57] and of developing automated pathways for characterization in this area[58–60].

Whether data collection is automated or manual, however, problem context dictates the performance of domain descriptors.

Macromolecular fingerprint representations

Generally, fingerprinting strategies involve converting molecular information into a numeric vector, such as a bit string, which expresses structural information. There are many different approaches to molecular fingerprinting, the most popular approaches can be broadly divided into three categories: substructure-keys-based fingerprints, topological or path-based fingerprints, and circular fingerprints[61]. Substructure-keys-based fingerprints set the bits of the associated bit string based on the presence or absence of a given chemical structure or features based upon a predefined dictionary[61]. Some popular examples of substructure-based fingerprints include MACCS, PubChem fingerprints, and Klekota–Roth fingerprints[62–64]. Both topological and circular fingerprints rely on a hashing function that characteristically abstracts molecular patterns from the macromolecular system into a vector[3]. Topological fingerprints such as Daylight fingerprints hash the connectivity between atoms up to a certain number of bonds[61], whereas circular fingerprints such as Extended-Connectivity fingerprints encode the chemical environment surrounding a given atom to a specified radius[65]. Thus, the attributes of these hashed descriptors are tailored to the system being modeled, rather than drawn from a predetermined schema. A principal challenge of adopting a hashed molecular fingerprinting-based chemical descriptor strategy is the connectivity of biopolymer subunits. Similarly, substructure-key fingerprints lack uniqueness as a function of structural arrangement. Some of this challenge may be alleviated by converting monomer unit fingerprints to those of their dimeric or oligomeric counterparts, though due to the stochastic nature of polymer species, the precise arrangement of subunits or substitutions (such as those in biopolymeric cellulose derivatives) is likely unknown[66]. One promising strategy to deal with this inherent ambiguity is to incorporate additional descriptors at various length scales (so-called hierarchical[67] or augmented scaled fingerprints[17], depending upon the included parameters) to more accurately describe a given polymer. For example, the success of the Polymer Genome project is underpinned by hierarchical fingerprints derived from data describing polymers at three length scales (atomic, molecular, morphological)[67]. Further, hierarchical fingerprinting has also shown success in biopolymer materials discovery, specifically in identifying naturally derived biopolymer candidates with improved thermomechanical and transport properties compared to existing synthetic materials[68].

It is additionally possible to combine a fingerprint with domain-specific descriptors, or graph descriptors, to correlate the molecular pattern vectors with relevant analytical and structural data. To this effect, experimentally-informed chemical fingerprints have been applied in tandem with macromolecular graph representations to perform polymer property prediction with graph neural networks[15,67]. A macromolecular fingerprinting approach should be selected with care, however, in tandem with a learning approach. A fingerprint benchmarking study compared the Morgan Fingerprint (MF), Molecular Embedding (ME), and Molecular Graph (MG) as alternative chemoinformatic descriptors under supervised learning, semi-supervised learning, and transfer learning schemes, with feed-forward neural networks predicting polymer density, melting temperature, and glass transition temperature[5]. Figure 3 depicts an exemplary fingerprint generation workflow, for each of the MF, ME, and MG, respectively. The study used 1442 homopolymer structures and labels available in the PolyInfo database, with two monomers of each structure modeled as training samples. While the ME had the best performance as a fingerprint, they found that the selected learning approach affected the explanatory chemical variables the model identified to map relationships between polymer attributes and properties with ME descriptors[5].

Fingerprinting offers simplicity of computation at some expense of interpretability, as the same fingerprint can be used to describe two different systems (referred to as a "bit collision")[5]. The allowance of bit collisions during fingerprinting prevents fingerprint representations from being applied directly for polymer inverse design. Additionally, latent spaces learned from fingerprint representations are considered "chemically incomplete" as the chemical features they encode are restricted to those contained by the dataset used during fingerprint hashing[14,16]. Latent space representations are typically created in deep representation learning, and apply architectures such as variational autoencoders for generative modeling tasks. In the context of biomaterials development, exploiting latent spaces holds much potential for increased performance. Latent spaces are well suited for global optimization as they are both continuous and differentiable[69,70] whereas complex chemical spaces tend to be challenging to optimize[71]. Some examples of latent space representations in polymeric materials include formulating novel biomaterial nanoaggregates of π-conjugated peptides[71] and developing polymers for extreme conditions[72].

As a whole, particularly in exploratory macromolecular informatics where inverse design is not a focus and chemical structure can be directly encoded, the simplicity and flexibility of fingerprinting can overcome trade-offs induced by bit collisions in view of ease of computation.

String representations

Models based upon string representations encode molecular structures and properties into strings in accordance with a predefined chemical knowledge framework. These strings are then treated as character sequences for featurization, akin to learning tasks in natural language processing. Hence, models that leverage string representations are popular for their facile interpretability, memory efficiency, and ready compatibility with natural language processing algorithms[3,7,38,73]. They are also generally considered chemically complete, which overcomes a limitation of macromolecular fingerprinting approaches, and are explicit in their description of polymer structures[14]. Generally, the task of implementing descriptor frameworks in language-based models involves first selecting a language-based encoding for chemical data, converting these encodings to distinct sub-units known as tokens, followed by the generation of characteristic embeddings using transformer architectures. The characteristic embeddings can then be applied as the input feature vector to the desired learning task.

A very popular encoding for string representation of polymers is the simplified molecular input line entry system (SMILES), which is natively designed for small molecules. Despite the fundamental differences between macromolecular and small molecule chemistries, SMILES representations have been used successfully in macromolecular machine learning. Two recent examples leveraging Bidirectional Encoder Representations from Transformers (BERT) architectures polyBERT[74] and TransPolymer[75] were both able to accurately perform polymer property prediction tasks by applying these language-based approaches with polymer SMILES strings as an input. Of note the connectivity problems as discussed earlier were resolved either by explicitly denoting connectivity in the associated SMILES string or through canonicalization. Other language-inspired models such as LSTM and n-gram type models have also been applied to biomacromolecular problem spaces such as the prediction of aggregate behaviors of polymers and biopolymers and predicting the radius of gyration[56,76]. Generally, the choice of architecture is imperative for the success of a given model based on string descriptors. While BERTs have shown tremendous success, including outperforming other more

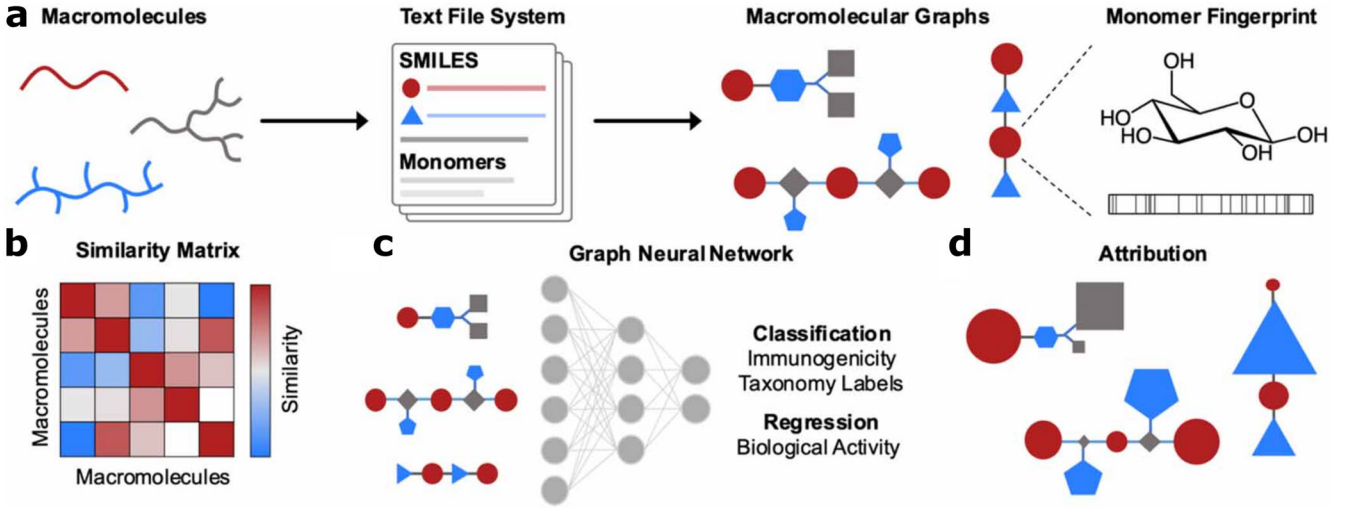

Fig. 4 Illustration of the generation and application of macromolecular graph representations for property prediction, cross-polymer comparison, and macromolecular interaction mechanism interpretation. Transformation path of raw macromolecular structures in the workflow, first converted into SMILES text files, then network graphs. **a** Graph nodes correspond to monomers, edges correspond to bonds, both of which are attributed to vectorized molecular fingerprints describing aspects of their underlying molecules. **b** Exemplary pair-wise similarity matrix derived from dimension-reduced representations of macromolecular species across the training library. **c** GNN computation of various interaction prediction labels from input macromolecular graphs. **d** Post-hoc graph attribution analysis explains underlying structures important to model-assigned interaction predictions. Adapted with permission from ref. [15] (copyright Somesh Mohapatra, Joyce An and Rafael Gòmez-Bombarelli, 2022).

traditional models (such as LSTMs) in polymer property prediction tasks[75] it is highly dependent on the system and modeling task.

Two more recently developed chemical string encoding frameworks are SELFIES and BIGSMILES. The SELFIES framework was developed primarily to overcome the limitations of SMILES for inverse design: every SELFIES representation corresponds to a chemically valid structure, which is not true of SMILES[52,53]. Both SELFIES and SMILES however are dedicated small molecule encodings and cannot uniquely encode polymer chemistry[7]. In light of this, the BIGSMILES framework was created for large polymer encodings. BIGSMILES can encode co-polymer information (homopolymer, random-, block co-polymer status), and distinguish linear, ring, and branched polymers[13].

Outside of preexisting string formats, physical analogs in biopolymer research can be represented as domain-specific string formats and leveraged as inputs to natural language processing workflows[3]. For example, prediction of the immunogenicity of glycans, a non-linear biological macromolecule, was achieved from deep learning models trained on 19,299 glycan examples, characteristically binned into "glycoword" monosaccharide groups[77].

It merits noting that string representations can also be applied to model biomacromolecules without the use of natural language models, similar to a fingerprint. For example, polymer functional groups encoded as SMILES can be converted to.mol files using the RDKit package to generate signature descriptors for directly training a machine learning model[1]. Using this methodology, one recent work identified monomer groups associated with macrophage-instructive behavior in meth(acrylate) and meth(acrylamide) polymers, specifically using supervised learning with multi-modal descriptors from high throughout co-polymer screening, optical microscopy, and SMILES representation[1].

With chemical string and natural language-derived descriptors, limitations are largely related to scaling strings from atomic to macromolecular scales. In the case of SMILES, as an explicit small-molecule representation, strings become too long to feasibly parse at polymer scale and do not account for the hierarchical and stochastic nature of polymer behavior[13]. However, more efficient abstractions of large molecule systems using BIGSMILES suffer from the opposite problem: the explicit organization of the sub-components within a polymer is lost in order to incorporate stochasticity in representation[14]. Finally, approaches that apply methodology from natural language processing face the same constraints that impact natural language research only in chemical context: extensive pretraining, data augmentation, and large dataset sizes are required for these approaches to succeed in macromolecular informatics, which can handicap progress where data is scarce in this field[7,78]. In these cases macromolecular fingerprints and graph representations, where the direct encoding of a chemical structure is feasible to avoid such pitfalls, are reported to outperform language models when provided the same amount of data[78]. Taken together, the suitability of applying chemical string and natural language-inspired descriptors in a macromolecular informatics project, over a simpler fingerprint or domain-specific descriptor, depends on several key factors. While these factors are contextually dependent upon the objectives of the work, they broadly include data availability, the acceptable macromolecule size resolution for representation and associated chemical space trade-offs, and limitations of the available computational resources.

In terms of labor-intensiveness, generating the initial string descriptors derived from a predefined knowledge frameworks such as SMILES is relatively straight-forward as it only requires knowledge of polymer structure. Pursuing a natural language processing workflow however requires additional familiarity with complex NLP domain concepts such as tokenization and transformer models, increasing the labor and computational background required. Alternatively, fingerprinting the SMILES representation with RDKit can more readily generate feature vectors from the string representation without requiring natural language processing background. Both such approaches are readily scaled computationally using only a priori structural knowledge, which in turn makes them less laborious than manual data collection or creating a de novo biomaterial domain-specific string representation.

Graph representations of macromolecules

In computer science, graphs are data structures constructed from a collection of nodes, typically depicted as circles, and the edges, depicted as lines, which indicate relationships between nodes. The Graph Neural Network (GNN) approach to deep learning became popularized as a tailwind of the deep learning renaissance brought on by Convolutional Neural Networks (CNN)[79]. CNNs learn multi-scale feature representations from Euclidean domains of data, and can be generalized on graphs through graph convolution operations. The ability to construct supervised mappings from non-Euclidean, graph-structured data through graph convolution represented a fundamental breakthrough in chemoinformatics and supervised learning. Accordingly, numerous structure-property prediction studies resulted from small molecule graphs, as well as studies of protein interface prediction[80–82]. There is significant overlap between the research aims in graph learning, and predictive design for macromolecules, given the abundance of networks in macromolecular bioinformatics[83]. The formidable breakthrough in protein folding prediction by the AlphaFold deep learning model trained using graph data representations is one exemplary success of what is possible at the intersection of these rapidly evolving fields[83,84].

The engine underlying success in graph structure-prediction tasks is known as representation learning. Specifically, representation learning refers to a workflow where training data is input to a GNN formatted as a graph, after which a GNN during training constructs its own vectorized encoding of the data by traversal of the input graphs[3]. The resulting vector (i.e., the "learned representation") can be applied similarly to a hashed fingerprint, as an input feature descriptor to a downstream task-specific predictive model such as a random forest, artificial neural network, etc[3]. Hence, the primary advantage of graph representation learning in macromolecular informatics is the compatibility afforded by the graph data structure (i.e., nodes and edges) with the physical organization of macromolecular materials as monomers with linkages and resulting topology. Representation-learned encodings, in addition to the basic attributes of a macromolecule found in a fingerprint, encode the specific syntax of the interconnections within the macromolecule being modeled[15]. As well, graph representation learning is not as sensitive to the size of the training set compared to natural language processing-based models, which improves the applicability of these methods for data-constrained research[78]. However, for researchers in macromolecular biomaterials design, selecting the precise nodes and edges to define for a GNN representation learning task, and the level of systemic abstraction they represent requires careful consideration in the context of the training dataset. Newly developed frameworks for modeling polymers as macromolecules for informatics can offer inspiration in this regard.

A chemistry-informed graph representation for macromolecules was recently developed, allowing for the quantification of structural similarity of 19,147 glycan biopolymers, in terms of both chemical and topological attributes, along with interpretable macromolecular supervised learning[15]. The representation applied for this is depicted in Fig. 4. Another recent work drew inspiration from polymer stochasticity to construct a graph representation framework for over 40,000 polymers as molecular ensembles, while incorporating chain architecture, monomer stoichiometry, and degree of polymerization in the descriptor set[78]. Additionally, a new framework has been released for end-to-end polymer informatics, PolyGrammar, which is the first to be chemically complete, molecularly explicit, physically valid, explainable, and invertible for generative polymer inverse design[14]. PolyGrammar is derived from a symbolic hypergraph representation and as proof of concept constructed representations for 600 polyurethane samples[14].

Broadly, advances in graph representation learning for macromolecular systems present clear opportunities for paradigm shifts in predictive model capabilities, as was observed with AlphaFold[84]. However, in macromolecular domains where there are challenges with data collection, data availability, or lacking standardization where training data is available, all present obstacles to the deployment of a universal framework for molecular representation at scale[7].

One approach to this would be using molecular descriptors to describe the sub-units of the polymeric biomaterial, and building a graph embedding from the set of sub-units. So long as the chemical space inherent in the sub-units is countable, the approach would scale. The challenge with this is defining scope to abstract a polymeric biomaterial into sub-units. Procedures for creating sub-unit definitions, for example coarse-grained polymer representations in physics-based simulations, are not always rigorously defined[56,85], and their selection will introduce a level of ambiguity in the predictive task. Along these lines, the application of this descriptor class can require some customization, which increases the labor intensiveness as compared to an off-the-shelf approach.

Future research at the intersection of graph representation learning and self-driving labs for the accelerated discovery of biomacromolecular materials will present a powerful opportunity to combine both standardized data availability and the best available macromolecular data representations, in real-time[86].

Outlook

There are many options to consider in selecting biomacromolecular machine learning descriptors for polymer interaction prediction tasks. As emphasized in this perspective, the objectives of the research project, access to data, labor-intensiveness of data collection, and downstream requirements for interpretability should be considered in balance at the outset of selecting or designing choice descriptors. The performance of a polymer interaction prediction model can be attributed to any one or combination of factors in data curation, feature engineering, or model design and training workflows. It is extremely challenging to identify polymer design space ranges that correspond to interaction behaviors of interest, and characterize those ranges reproducibly at scale, with measurement error that does not obfuscate the desired signal. Knowing this, a "one-size fits all" approach to feature selection and modeling across interaction prediction tasks for polymeric biomaterials is unlikely to accommodate all varieties of relevant domain-specific factors. The heterogeneous nature of data curation in the field underpins the wide variety of feature engineering methodologies discussed in this perspective: domain-specific, fingerprint, string, and graph descriptors.

In navigating biopolymeric feature selection, where exploratory proof of concept work is the focus, employing descriptors that are both simple and interpretable establishes trust in the data and builds scientific intuition. Alternatively, in scaling a model for deployment with an active learning system or accelerated materials discovery platform, descriptors that support model generalization and capacity for inverse design can be imperative. In either case, the process of feature engineering for a predictive task is often iterative, and ultimately reflective of the inherent predictive power of the independent variables being used to describe the target. Along these lines, there has been a shift at the forefront of machine learning research from "model-centric AI" to "data-centric AI," which reflects the growing recognition across AI-accelerated research domains that efforts to improve the quality of training data can be more productive than efforts focused on model optimization for specific tasks[87]. In terms of data-centricity, it is hard to beat incremental value brought forward by applying high-resolution analytical tools such as mass spectroscopy and

nuclear magnetic resonance to enrich descriptor quality, and thus model quality. In opposition, it is very time-consuming to apply traditional analytical approaches for data generation at scale, which motivates the development of automated, scalable, data collection approaches. While one size does not fit all today, continued research efforts to automate high-resolution biomacromolecular data collection, and accurately encode biomacromolecular interaction phenomena as features, are expected to enable the next generation of predictive biomaterial polymer designs.

DATA AVAILABILITY

The data presented in this work are available at the referenced original sources.

Received: 7 November 2022; Accepted: 8 May 2023;
Published online: 12 June 2023

REFERENCES

1. Rostam, H. M. et al. Immune-instructive polymers control macrophage phenotype and modulate the foreign body response. *In Vivo Matter* **2**, 1564–1581 (2020).
2. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
3. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **12**, 1–22 (2020).
4. Fernández-Torras, A., Comajuncosa-Creus, A., Duran-Frigola, M. & Aloy, P. Connecting chemistry and biology through molecular descriptors. *Curr. Opin. Chem. Biol.* **66**, 102090 (2022).
5. Ma, R., Liu, Z., Zhang, Q., Liu, Z. & Luo, T. Evaluating polymer representations via quantifying structure-property relationships. *J. Chem. Inform. Model* **59**, 3110–3119 (2019).
6. Jones, D. E., Ghandehari, H. & Facelli, J. C. A review of the applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles. *Comput Methods Prog. Biomed.* **132**, 93–103 (2016).
7. Kumar, R. Materiomically designed polymeric vehicles for nucleic acids: quo vadis? *ACS Appl. Bio Mater.* **5**, 2507–2535 (2022).
8. Upadhya, R. et al. Automation and data-driven design of polymer therapeutics. *Adv. Drug Deliv. Rev.* **171**, 1–28 (2021).
9. Cencer, M. M., Moore, J. S. & Assary, R. S. Machine learning for polymeric materials: an introduction. *Polym. Int.* **71**, 537–542 (2022).
10. Cravero, F. et al. Feature selection for polymer informatics: evaluating scalability and robustness of the FS4RVDD algorithm using synthetic polydisperse data sets. *J. Chem. Inform. Model* **60**, 592–603 (2020).
11. Kumar, R. et al. Efficient polymer-mediated delivery of gene-editing ribonucleoprotein payloads through combinatorial design, parallelized experimentation, and machine learning. *ACS Nano* **14**, 17626–17639 (2020).
12. Watchorn, J. et al. Untangling mucosal drug delivery: engineering, designing, and testing nanoparticles to overcome the mucus barrier. *ACS Biomater. Sci. Eng.* **8**, 1396–1426 (2022).
13. Lin, T. S. et al. BigSMILES: a structurally-based line notation for describing macromolecules. *ACS Cent. Sci.* **5**, 1523–1531 (2019).
14. Guo, M. et al. Polygrammar: grammar for digital polymer representation and generation. *Adv. Sci.* **9**, 2101864 (2022).
15. Mohapatra, S., An, J. & Gómez-Bombarelli, R. Chemistry-informed macromolecule graph representation for similarity computation, unsupervised and supervised learning. *Mach. Learn Sci. Technol.* **3**, 015028 (2022).
16. Xu, P., Chen, H., Li, M. & Lu, W. New opportunity: machine learning for polymer materials design and discovery. *Adv. Theory Simul.* **5**, 2100565 (2022).
17. Patel, R. A., Borca, C. H. & Webb, M. A. Featurization strategies for polymer sequence or composition design by machine learning. *Mol. Syst. Des. Eng.* **7**, 661–676 (2022).
18. Richardson, J. J. & Caruso, F. Nanomedicine toward 2040. *Nano Lett.* **20**, 1481–1482 (2020).
19. Germain, M. et al. Delivering the power of nanomedicine to patients today. *J. Control. Release* **326**, 164–171 (2020).
20. Fadeel, B. & Alexiou, C. Brave new world revisited: focus on nanomedicine. *Biochem. Biophys. Res. Commun.* **533**, 36–49 (2020).
21. Serov, N. & Vinogradov, V. Artificial intelligence to bring nanomedicine to life. *Adv. Drug Deliv. Rev.* **184**, 114194 (2022).
22. Meyer, T. A., Ramirez, C., Tamasi, M. J. & Gormley, A. J. A user's guide to machine learning for polymeric biomaterials. *ACS Polym. Au.* **3**, 141–157 (2023).
23. Lazarovits, J. et al. Supervised learning and mass spectrometry predicts the in vivo fate of nanomaterials. *ACS Nano* **13**, 8023–8034 (2019).
24. Bannigan, P. et al. Machine learning directed drug formulation development. *Adv. Drug Deliv. Rev.* **175**, 113806 (2021).
25. Kerner, J., Dogan, A. & Von Recum, H. Machine learning and big data provide crucial insight for future biomaterials discovery and research. *Acta Biomater.* **130**, 54–65 (2021).
26. Friederich, P., Krenn, M., Tamblyn, I. & Aspuru-Guzik, A. Scientific intuition inspired by machine learning-generated hypotheses. *Mach. Learn Sci. Technol.* **2**, 025027 (2021).
27. Xu, Y. et al. Deep dive into machine learning models for protein engineering. *J. Chem. Inform. Model* **60**, 2773–2790 (2020).
28. Kwaria, R. J., Mondarte, E. A. Q., Tahara, H., Chang, R. & Hayashi, T. Data-driven prediction of protein adsorption on self-assembled monolayers toward material screening and design. *ACS Biomater. Sci. Eng.* **6**, 4949–4956 (2020).
29. Le, T. C., Penna, M., Winkler, D. A. & Yarovsky, I. Quantitative design rules for protein-resistant surface coatings using machine learning. *Sci. Rep.* **9**, 265 (2019).
30. Kumar, J. N. et al. Machine learning enables polymer cloud-point engineering via inverse design. *NPJ Comput. Mater.* **5**, 73 (2019).
31. Wang, A. Y.-T. et al. Machine learning for materials scientists: an introductory guide toward best practices. *Chem. Mater.* **32**, 4954–4965 (2020).
32. Lössl, P., Waterbeemd, M. & Heck, A. J. The diverse and expanding role of mass spectrometry in structural and molecular biology. *EMBO J.* **35**, 2634–2657 (2016).
33. Bouwmeester, R., Gabriels, R., Van Den Bossche, T., Martens, L. & Degroeve, S. The age of data-driven proteomics: how machine learning enables novel workflows. *Proteomics* **20**, 1900351 (2020).
34. Corbo, C. et al. Analysis of the human plasma proteome using multi-nanoparticle protein corona for detection of Alzheimer's disease. *Adv. Health. Mater.* **10**, 2000948 (2021).
35. Willcox, K. E., Ghattas, O. & Heimbach, P. The imperative of physics-based modeling and inverse theory in computational science. *Nat. Comput Sci.* **1**, 166–168 (2021).
36. Marchetti, R. et al. "Rules of Engagement" of protein-glycoconjugate interactions: a molecular view achievable by using NMR spectroscopy and molecular modeling. *ChemistryOpen* **5**, 274–296 (2016).
37. Moradi Kashkooli, F., Soltani, M., Souri, M., Meaney, C. & Kohandel, M. Nexus between in silico and in vivo models to enhance clinical translation of nanomedicine. *Nano Today* **36**, 101057 (2021).
38. Sanchez-Lengeling, B. et al. A Bayesian approach to predict solubility parameters. *Adv. Theory Simul.* **2**, 1800069 (2019).
39. Erlebach, A. et al. Predicting solubility of small molecules in macromolecular compounds for nanomedicine application from atomistic simulations. *Adv. Theory Simul.* **3**, 2000001 (2020).
40. Jackson, N. E. Coarse-graining organic semiconductors: the path to multiscale design. *J. Phys. Chem. B* **125**, 485–496 (2021).
41. Dhamankar, S. & Webb, M. A. Chemically specific coarse-graining of polymers: methods and prospects. *J. Polym. Sci.* **59**, 2613–2643 (2021).
42. Liang, H., Webb, M. A., Chawathe, M., Bendejacq, D., & De Pablo, J. J. Understanding the structure and rheology of galactomannan solutions with coarse-grained modeling. *Macromolecules* **56**, 177–187 (2022).
43. Watchorn, J., Burns, D., Stuart, S. & Gu, F. X. Investigating the molecular mechanism of protein–polymer binding with direct saturation compensated nuclear magnetic resonance. *Biomacromolecules* **23**, 67–76 (2022).
44. Madiona, R. M. T., Winkler, D. A., Muir, B. W. & Pigram, P. J. Optimal machine learning models for robust materials classification using ToF-SIMS data. *Appl. Surf. Sci.* **487**, 773–783 (2019).
45. Watchorn, J., Stuart, S., Burns, D. C. & Gu, F. X. Mechanistic influence of polymer species, molecular weight, and functionalization on mucin–polymer binding interactions. *ACS Appl. Polym. Mater.* **4**, 7537–7546 (2022).
46. Fino, R. et al. Introducing the CSP analyzer: a novel machine learning-based application for automated analysis of two-dimensional NMR spectra in NMR fragment-based screening. *Comput. Struct. Biotechnol. J.* **18**, 603–611 (2020).
47. Tamasi, M. J. et al. Machine learning on a robotic platform for the design of polymer–protein hybrids. *Adv. Mater.* **34**, 2201809 (2022).
48. Shan, X. et al. Synthesis and evaluation of methacrylated poly(2-ethyl-2-oxazoline) as a mucoadhesive polymer for nasal. *Drug Deliv. ACS Appl. Polym. Mater.* **3**, 5882–5892 (2021).
49. Khutoryanskiy, V. V. Beyond PEGylation: alternative surface-modification of nanoparticles with mucus-inert biomaterials. *Adv. Drug Deliv. Rev.* **124**, 140–149 (2018).
50. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **92**, 014106 (2015).
51. Park, N. H. et al. A recommender system for inverse design of polycarbonates and polyesters. *Macromolecules* **53**, 10847–10854 (2020).

52. Nigam, A. et al. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chem. Sci.* **12**, 7079–7090 (2021).
53. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn Sci. Technol.* **1**, 045024 (2020).
54. Singh, A. V. et al. Artificial intelligence and machine learning empower advanced biomedical material design to toxicity prediction. *Adv. Intell. Syst.* **2**, 2000084 (2020).
55. Ma, S. & Dai, Y. Principal component analysis based methods in bioinformatics studies. *Brief Bioinform* **12**, 714–722 (2011).
56. Webb, M. A., Jackson, N. E., Gil, P. S. & de Pablo, J. J. Targeted sequence design within the coarse-grained polymer genome. *Sci. Adv.* **6**, eabc6216 (2020).
57. Gormley, A. J. & Webb, M. A. Machine learning in combinatorial polymer chemistry. *Nat. Rev. Mater.* **6**, 642–644 (2021).
58. Mohapatra, S. et al. Deep learning for prediction and optimization of fast-flow peptide synthesis. *ACS Cent. Sci.* **6**, 2277–2286 (2020).
59. Leibfarth, F. A., Johnson, J. A. & Jamison, T. F. Scalable synthesis of sequence-defined, unimolecular macromolecules by Flow-IEG. *Proc. Natl Acad. Sci.* **112**, 10617–10622 (2015).
60. Tamasi, M., Kosuri, S., DiStefano, J., Chapman, R. & Gormley, A. J. Automation of controlled/living radical polymerization. *Adv. Intell. Syst.* **2**, 1900126 (2020).
61. Cereto-Massagué, A. et al. Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).
62. Klekota, J. & Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **24**, 2518–2525 (2008).
63. Bolton, E. E., Wang, Y., Thiessen, P. A. & Bryant, S. H. PubChem: integrated platform of small molecules and biological activities. In *Proc. Annual Reports in Computational Chemistry* (eds. Wheeler, R. A. & Spellmeyer, D. C.) 217–241 (Elsevier, 2008). https://doi.org/10.1016/S1574-1400(08)00012-1.
64. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inform. Comput. Sci.* **42**, 1273–1280 (2002).
65. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inform. Model.* **50**, 742–754 (2010).
66. Patel, R. A. & Webb, M. A. Data-driven design of polymer-based biomaterials: high-throughput simulation, experimentation, and machine learning. *ACS Appl. Bio Mater.* https://doi.org/10.1021/acsabm.2c00962 (2023).
67. Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer genome: a data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C* **122**, 17575–17585 (2018).
68. Kuenneth, C. et al. Bioplastic design using multitask deep neural networks. *Commun. Mater.* **3**, 96 (2022).
69. Calandra, R., Peters, J., Rasmussen, C. E. & Deisenroth, M. P. Manifold Gaussian processes for regression. In *Proc. International Joint Conference on Neural Networks (IJCNN)* 3338–3345 (IEEE, 2016). https://doi.org/10.1109/IJCNN.2016.7727626.
70. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
71. Shmilovich, K. et al. Discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation. *J. Phys. Chem. B* **124**, 3873–3891 (2020).
72. Batra, R. et al. Polymers for extreme conditions designed using syntax-directed variational autoencoders. *Chem. Mater.* **32**, 10489–10500 (2020).
73. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Model* **28**, 31–36 (1988).
74. Kuenneth, C. & Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. Preprint at http://arxiv.org/abs/2209.14803 (2022).
75. Xu, C., Wang, Y. & Barati Farimani, A. TransPolymer: a transformer-based language model for polymer property predictions. *npj Comput. Mater.* **9**, 64 (2023).
76. Bhattacharya, D., Kleeblatt, D. C., Statt, A. & Reinhart, W. F. Predicting aggregate morphology of sequence-defined macromolecules with recurrent neural networks. *Soft Matter* **18**, 5037–5051 (2022).
77. Bojar, D., Powers, R. K., Camacho, D. M. & Collins, J. J. Deep-learning resources for studying glycan-mediated host-microbe interactions. *Cell Host Microbe* **29**, 132–144.e3 (2021).
78. Aldeghi, M. & Coley, C. W. A graph representation of molecular ensembles for polymer property prediction. *Chem. Sci.* **13**, 10486–10498 (2022).
79. Zhou, J. et al. Graph neural networks: a review of methods and applications. *AI Open* **1**, 57–81 (2020).
80. Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S. & Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inform. Model* **57**, 1757–1772 (2017).
81. Fout, A., Byrd, J., Shariat, B. & Ben-Hur, A. Protein interface prediction using graph convolutional networks. In *Proc. 31st Conference Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) **30**, 6530–6539 (2017).
82. Duvenaud, D. K. et al. Convolutional networks on graphs for learning molecular fingerprints. In: *Proc. Advances in Neural Information Processing Systems* (eds. Cortes, C., Lawrence, N., Lee, D., Sugiyama, M. & Garnett, R.) **28**, 2224–2232 (2015).
83. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. Preprint at http://arxiv.org/abs/2010.09885 (2020).
84. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
85. Webb, M. A., Delannoy, J.-Y. & de Pablo, J. J. Graph-based approach to systematic molecular coarse-graining. *J. Chem. Theory Comput* **15**, 1199–1208 (2019).
86. Seifrid, M., Hattrick-Simpers, J., Aspuru-Guzik, A., Kalil, T. & Cranford, S. Reaching critical MASS: crowdsourcing designs for the next generation of materials acceleration platforms. *Matter* **5**, 1972–1976 (2022).
87. Eyuboglu, S., Karlaš, B., Ré, C., Zhang, C. & Zou, J. dcbench: a benchmark for data-centric AI systems. In *Proc. Sixth Workshop on Data Management for End-To-End Machine Learning* 1–4 (ACM, 2022). https://doi.org/10.1145/3533028.3533310.

ACKNOWLEDGEMENTS

This work was supported by NSERC Discovery Grant #06441 and the NSERC Senior Industrial Research Chair program. S.S. is supported by the NSERC Alexander Graham Bell Canada Graduate Scholarship and the Canadian Federation of University Women 1989 École Polytechnique Commemorative Award. J.W. is supported by Queen Elizabeth II/Dupont Canada Scholarship in Science and Technology and the Mclean Foundation Graduate Scholarships in Science And Technology.

AUTHOR CONTRIBUTIONS

S.S. and J.W. conceptualized the article. All authors contributed to writing and editing the article with emphasis on the contributions by S.S. All figures were adapted or created by J.W. F.X.G. supervised the work.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Frank X. Gu.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.