






Cite this: *Chem. Commun.*, 2023, 59, 14197

# Frontiers in nonviral delivery of small molecule and genetic drugs, driven by polymer chemistry and machine learning for materials informatics

Jeffrey M. Ting, \* Teresa Tamayo-Mendoza, Shannon R. Petersen, Jared Van Reet,  Usman Ali Ahmed, Nathaniel J. Snell,  John D. Fisher, Mitchell Stern and Felipe Oviedo

Materials informatics (MI) has immense potential to accelerate the pace of innovation and new product development in biotechnology. Close collaborations between skilled physical and life scientists with data scientists are being established in pursuit of leveraging MI tools in automation and artificial intelligence (AI) to predict material properties *in vitro* and *in vivo*. However, the scarcity of large, standardized, and labeled materials data for connecting structure–function relationships represents one of the largest hurdles to overcome. In this Highlight, focus is brought to emerging developments in polymer-based therapeutic delivery platforms, where teams generate large experimental datasets around specific therapeutics and successfully establish a design-to-deployment cycle of specialized nanocarriers. Three select collaborations demonstrate how custom-built polymers protect and deliver small molecules, nucleic acids, and proteins, representing ideal use-cases for machine learning to understand how molecular-level interactions impact drug stabilization and release. We conclude with our perspectives on how MI innovations in automation efficiencies and digitalization of data—coupled with fundamental insight and creativity from the polymer science community—can accelerate translation of more gene therapies into lifesaving medicines.

Received 22nd September 2023,  
Accepted 2nd November 2023

DOI: 10.1039/d3cc04705a

rsc.li/chemcomm

Nanite, Inc., Boston, Massachusetts 02109, USA. E-mail: jeff@nanitebio.com



Jeffrey M. Ting

*Jeff Ting is a Senior Scientist at Nanite. His research focuses on generating large-scale classes of polymer nanoparticles, equipped with precisely tailored chemical and physical attributes for controlled storage/release, to enable high-throughput in vitro–in vivo screening of gene delivery and AI-driven materials design. Previously, he worked in the 3M Corporate Research Materials Lab's Materials Informatics Group. He received his PhD in Chemical Engineering (University of Minnesota, 2016) and worked as a NIST-CHiMaD Postdoctoral Fellow at the Pritzker School of Molecular Engineering (University of Chicago, 2019) under the Center for Hierarchical Materials Design, supported by NIST and the Materials Genome Initiative.*

## 1. Introduction

Polymers have entered every aspect of society, from familiar consumer products to high-value technological applications in electronics, energy, and healthcare.<sup>1,2</sup> In medicine, polymers have found particular utility in the development of early controlled drug delivery systems for oral and dermal medications—prototypical of such systems are acrylics and methacrylics, carbohydrates and, more recently, resorbable systems such as polyglycolides.<sup>3</sup> The rapid embrace of macromolecular therapeutic entities (including antibodies, proteins, and gene therapies) now offers both great challenges and great opportunities to polymer chemistry. Genetic drugs in particular appear especially poised to be transformative, driven by the growing understanding of the biological underpinnings of genetic disease and immunology. In tandem with this growth there is a recognized need to develop more robust nonviral nanocarriers that achieve delivery with high specificity, reliability, and safety. However, there are as of yet, few polymer-based vectors approved for genetic therapeutics delivery.

Unlike their small molecule counterparts, therapeutic biologics present distinct challenges. DNA, RNA, and genomic editing ribonucleoproteins<sup>4</sup> are larger, hydrophilic, ionic, and

## Highlight

prone to degradation. Prospective polymer delivery systems need to balance opposing attributes for these payloads by providing (i) colloidal stabilization across multiple biological barriers,<sup>5</sup> and (ii) efficient payload release at the site of action.<sup>6</sup> This dichotomy complicates the (mostly) well-understood molecular engineering approaches used for small molecule drugs that rely on conventional controlled drug delivery principles and computational foundations.

Because of the vast design space of chemistries and architectures, it remains difficult to intuitively devise an ideal polymer vector that can fulfill every desired function in macromolecular biologics delivery. Nevertheless, polymer chemistry has advanced to the point where unlimited structures can be created, as described in recent perspectives on controlled reversible-deactivation radical polymerization,<sup>7</sup> chemical functionalization,<sup>8</sup> site-specific bioconjugation,<sup>9</sup> and electrostatic self-assembly.<sup>10</sup> High-throughput synthesis and screening campaigns have taken advantage of this versatility to tailor specialized polymers around a single drug of interest.<sup>11–13</sup> Challenges remain, however, in the efficient deployment of the vast toolbox of potential polymeric delivery systems across an enormously divergent set of therapeutic modalities.

One potential solution to navigate this immense design space is the marriage of experimental and synthetic data with materials informatics (MI) to develop a deeper understanding of structure–function relationships between polymer-mediated binding and delivery of various drugs. MI depends on collecting, cleaning, and organizing machine actionable data into a framework to leverage machine learning (ML) algorithms and artificial intelligence (AI) applications.<sup>14,15</sup> Unfortunately, materials data curation is often a formidable challenge because information sources are dispersed, inhomogeneous, and inaccessible. This challenge is particularly true for polymer science, where progress has lagged on laying the groundwork for reconciling large polymer datasets with digitalization.<sup>16–19</sup>

In this short Highlight article, we feature three examples that apply these principles and demonstrate polymer synthesis/screening campaigns for three distinct cargos: (1) small molecule drugs, (2) nucleic acids, and (3) proteins. These vignettes show how rapid data generation can facilitate ML models to produce multifunctional nanoparticle candidates for the therapeutic of interest (Fig. 1). High-throughput polymer chemistry, nonviral drug delivery, and MI are connected through close collaborations across multiple teams with distinct skillsets in each use case. A glossary of MI terms and methodologies is provided at the end of this Highlight for readers' reference. Finally, we provide an outlook for expanding these themes to pharmaceutical applications in nonviral gene therapy. The breadth and diversity of genetic drugs span physiochemical attributes that must be accounted for in data-driven polymer design from a near infinite chemical space. To this end, laboratory workflow automation and data management best practices are discussed that can prioritize therapeutic formulations with higher likelihood of successful delivery. In our view, assembling these physical and digital pieces together can usher in the next era of potent, affordable genetic drugs to market.

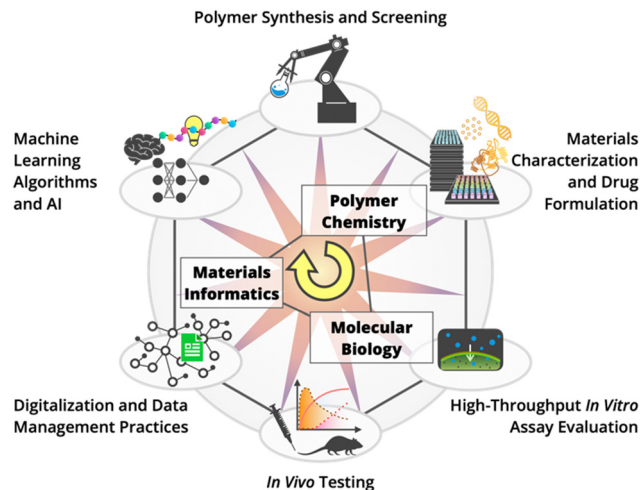


Fig. 1 Illustration of an emerging paradigm for nonviral drug delivery development, integrating elements of polymer chemistry, molecular biology, and materials informatics. This platform requires knowledge and skillsets from polymer synthesis/screening and materials characterization, high-throughput biological testing of therapeutic efficacy/safety, and data science and engineering using machine actionable data.

## 2. Polymer design propelled by MI

### 2.1. Small molecule drugs

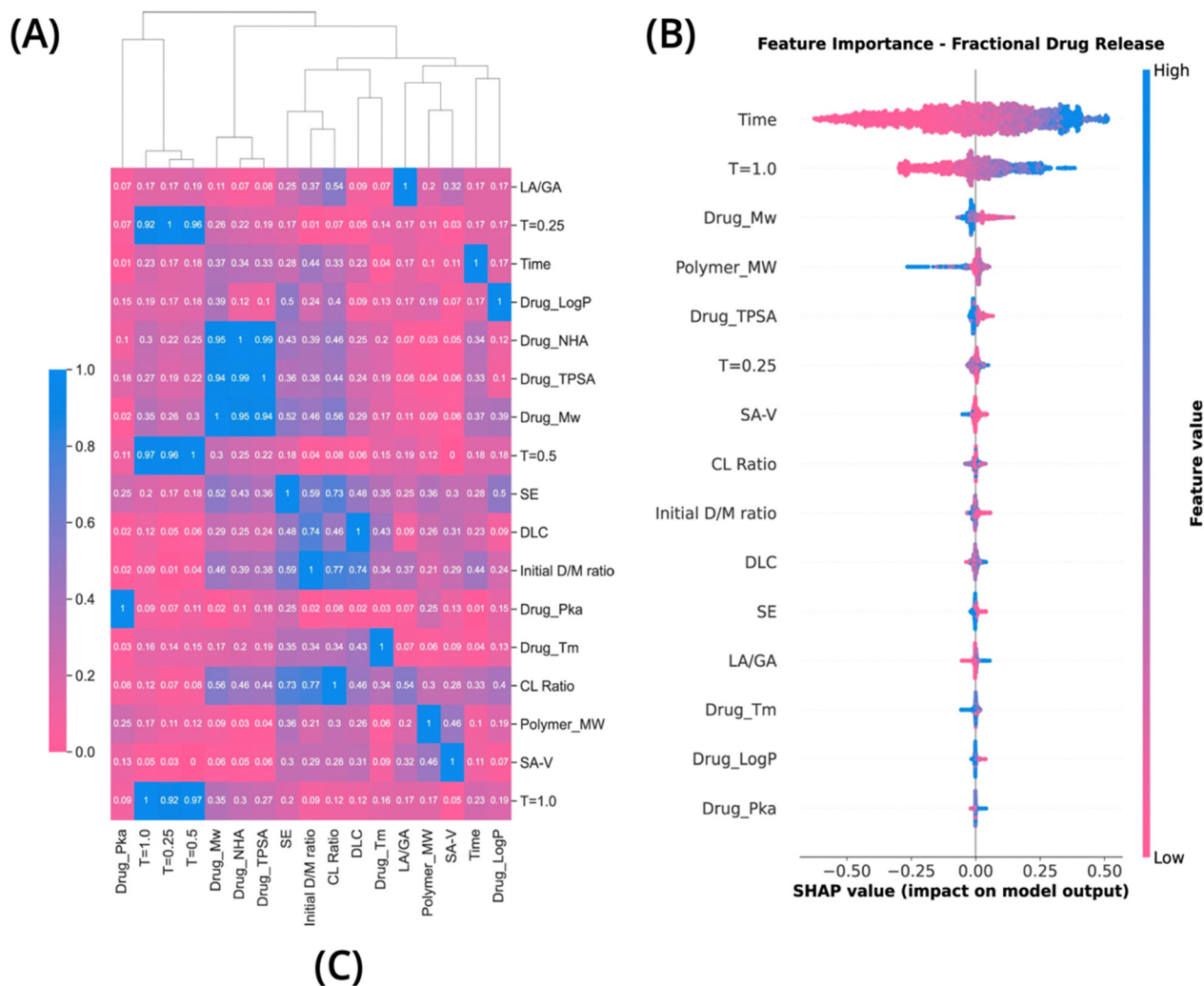
Polymer-based long-acting injectables (LAIs) are drug formulations designed to prolong the stability and bioavailability of a therapeutic by controlling and sustaining its release. Clinical translation of LAIs has long been limited by the relatively few polymeric families approved for parenteral administration. Poly(lactide-co-glycolide) (PLGA) is a biodegradable polymer and the key excipient for the majority of the ~30 clinically approved LAI products.<sup>20</sup> Thus, PLGA-based systems remain a widely studied platform for satisfying the downstream regulatory process. While this has limited exploration in chemical diversity for polymeric LAIs, there is abundant formulation data within the PLGA literature on drug stability, loading capacity, and *in vitro* release profiles that can be extracted from experimental and simulation literature<sup>21</sup> and can, in theory, be used to design novel polymers for parenteral administration. While these reports provide insight to develop LAI systems, they require a considerable amount of experimental time or are computationally intensive.

Banningan *et al.*<sup>22</sup> recently explored a ML approach to predict fractional drug release and proposed a universal framework for designing LAI systems. They curated a data set of 43 drug–polymer combinations that consist of commercially available polymers with 31 783 partial and 181 complete drug release properties from previous publications. As a starting point, 17 descriptors from experimental conditions and physicochemical properties of the drug and polymers were examined. After training different models and assessing the performance of release predictions, the authors selected a tree-based regression model called light gradient-boosting machine (LGBM) for further refinement. Two models were trained: the first excluded any

points of the drug release curves in the training data features, while the second used three initial measurements included as the features. The team selected the second model for further analysis.

Agglomerative hierarchical clustering based on Spearman's rank correlation was performed to remove redundant variables from the final predictive model (Fig. 2(A)). This statistical test determines the presence of the monotonic relationship between two variables. By arranging variables into hierarchy of clusters from this test, they found that removing two features from clusters with strong correlations (*i.e.*, the fractional drug

release at 0.5 day or  $T = 0.5$  and the number of heteroatoms or NHA) resulted in a model with similar accuracy. Meanwhile, despite a strong correlation between drug molecular weight (MW) and topological polar surface area (TPSA), the removal of these features reduced model accuracy. This example shows how descriptors like  $T = 0.5$  and NHA were removed while others like MW and TPSA were retained, resulting in 15 finalized features. To further determine which features are important in the model, a SHapley Additive exPlanations (SHAP) analysis was performed (Fig. 2(B)). SHAP is a method to explain predictive



**Fig. 2** (A) Heatmap of the absolute Spearman's rank correlation of the initial 17 input features for LAI development. The dendrogram displays agglomerative hierarchical clustering analysis, *e.g.*,  $T = 0.25$ ,  $T = 1.0$ , and  $T = 0.5$ . Pink and blue represent 0.0 and 1.0 correlation values, respectively. (B) Swarm plot of SHAP values of the 15-feature LGBM model. The colors pink and blue represent relatively low and high values, respectively. (C) Table with the proposed design criteria to select "fast" and "slow" drug-release profiles based on the 15-feature LGBM model, SHAP analysis, and observed trends in PCA and tSNE plots. For instance, low molecular weights of drug cargo and polymer system are associated with a "fast" drug release profile. Adapted with permission from ref. 22. Copyright 2023 Nature Publishing Group.

outputs of ML models by computing the relative contribution of each input feature. It was found that the model's most influential features were time, and specifically  $T = 1.0$ , the drug's one day measurement of release. Other significant drivers were the MWs of the polymers and drugs, respectively. The results did not display the potential synergy of other features, as suggested by the analysis of feature removal by agglomerative hierarchical clustering analysis.

Finally, the authors proposed and experimentally tested two LAI formulations focused on microparticles that are easy to produce based on commercially available PLGAs. They used the experimental measurement  $T = 1.0$  as a proxy, one of the most influential features in the SHAP analysis. They suggested that the low-values of the fractional drug release might correspond to "fast" sustained release profiles compared to LAI systems with a relatively high value ("slow" release). Moreover, by analyzing low-dimensional clustered plots (principal component analysis; PCA) and an unsupervised clustering algorithm (T-distributed Stochastic Neighbor Embedding; t-SNE), it was observed that some features were generally related to the values of fractional drug released at  $T = 1.0$  and, therefore, to a "slow" or "fast" release. Furthermore, they proposed a LAI design criterion (Fig. 2(C)) and selected two drug-PLGA pairs to function as "fast" and "slow" release systems. For the first LAI, a 10 kDa PLGA and salicylic acid (SA) were chosen for their relatively low MWs, relatively low  $\log P$  of SA, and relatively low TPSA value of SA. By comparison, a "slow" release LAI system consisted of a 50 kDa PLGA and olaparib (OLA), where both components have relatively high MWs, relatively high  $\log P$  of OLA, and high TPSA value of OLA. They prepared and characterized samples using an oil-in-water emulsion method<sup>23</sup> and observed excellent agreement between predicted and experimental release profiles. The authors speculate that further improvements could be effected by incorporating factors that were excluded from the model, such as polymer degradation in the PLGA formulation. Nevertheless, this benchmarks a powerful method to establish design rules for other LAI pairings, assuming that such prospective systems have access to sufficient training data.

## 2.2. Nucleic acid and ribonucleoprotein delivery

In the ~40 years history of gene therapy, the field has been dominated by viral vectors and, more recently, by lipid nanoparticles, with both modalities achieving significant successes and limitations. A relatively unexplored avenue for nucleic acid delivery has been the use of polymers that are capable of forming complexes (polyplexes) with nucleic acid cargos. Generally speaking, polyplexes are positively charged polymeric vectors that bind to cargos and form nanoparticles with sizes in order of approximately 100 nm. In order to effectively transfect cells, polyplexes rely on a delicate balance of chemical and physical features. It is important to recognize that polyplexes are dynamic nanoparticles—that is, because they self-assemble at the molecular level from non-covalent interactions with energies of multiples of  $kT$ ,<sup>24</sup> these nanoscale assemblies can rearrange under different environmental conditions. While numerous studies have investigated the optimization of commercially available polyethylenimine (PEI) and its variants, a one-size-fits-all

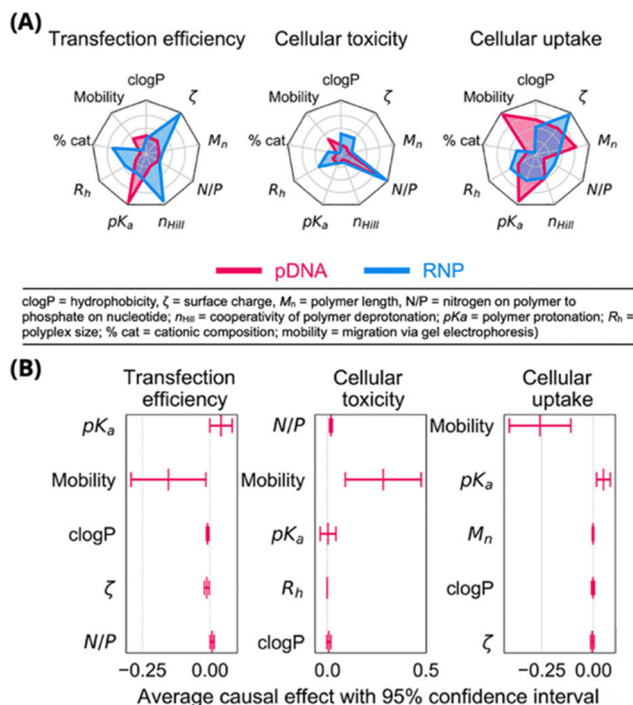
approach for different exogenous nucleic acid cargo often cannot be utilized. What has been lacking is an appreciation of a design rule approach for polymers that can be readily tailored to deliver specific biological cargos to desired extracellular and intracellular destinations.

Recently, Kumar *et al.* addressed this issue by revisiting an established polymer library and determining if the same design constraints apply for delivery of a different cargo.<sup>25</sup> These 43 polymers spanned commonly investigated cationic and hydrophilic monomers and were originally investigated as vectors for ribonucleoprotein (RNP) delivery. This study demonstrated that successful delivery of RNP cargo was most dependent on the polyplex surface charge and the degree of cooperativity during polymer deprotonation ( $n_{\text{Hill}}$ ). In the new study, the same library was re-examined with the following objectives: (1) identify polymers that efficiently facilitate intracellular delivery of plasmid DNA (pDNA), (2) determine if design constraints applied for RNP payloads are relevant to pDNA payloads, (3) co-deliver RNP and pDNA payloads for homology-directed repair, and (4) translate the results to specific targets. Eight candidates showed substantial increase in transgene expression, with the polymer P38 (comprising 2-(diisopropylamino)ethyl methacrylate and 2-hydroxyethyl methacrylate monomers, poly(DIPAEMA<sub>52</sub>-*st*-HEMA<sub>50</sub>)) as the lead candidate from the library screen. It was determined by quantitative confocal microscopy that P38 was able to effectively deliver pDNA to two distinct cell types, HEK293T and ARPE-19, showing relatively high levels of nuclear import and the ability to escape endosomal compartments.

P38 was also determined to be the lead candidate for RNP delivery, and it was initially suspected that the polymeric design criteria might be identical for both RNP and pDNA delivery. Thus, to further elucidate any structure–function relationships between polymer attributes and payload type, SHAP analysis was used. The SHAP analysis revealed that the design parameters affecting cellular uptake, delivery efficiency, and toxicity are all cargo dependent (Fig. 3). Notably, RNP delivery is dependent on hydrophobic interactions in addition to electrostatic interactions, and that these are both necessary for cytosolic release. On the other hand, hydrophobic interactions are negligible for successful pDNA delivery, which relies on the optimization of polycation protonation equilibria and pDNA binding affinity. Despite the payload dependent divergence in vector design, it is important to note that polymer compositions such as P38 can simultaneously satisfy the requirements of both payloads. This was demonstrated by using P38 to successfully co-deliver RNP and pDNA payloads for homology driven repair at a higher rate than JetPEI, a commercial polymer routinely used as a gold standard in gene delivery. In addition to identifying a promising polymer for delivery of two distinct payloads, this important work introduces a robust framework for deconvoluting payload-specific structure–function relationships.

## 2.3. Polymer–protein hybrids

One of the most advantageous characteristics of polymer nanoparticles over alternatives is the potential for intracellular



**Fig. 3** (A) SHAP illustrates the importance and contributions of polyplex features to delivery efficiency, cellular toxicity, and uptake for pDNA (pink) and RNP (blue) payloads. (B) Average treatment effect (ATE) analysis estimates causal structure–function trends of pDNA polyplexes from the top features from SHAP analysis. Positive and negative effects (error bars denote 95% CI) denote antagonistic and antagonistic relationships, respectively. Adapted with permission from ref. 26. Copyright 2022 American Chemical Society.

delivery of biomacromolecules, such as large RNP or antibody payloads. Indeed, direct cytosolic delivery of proteins has recently been demonstrated for a handful of engineered polymer systems.<sup>26–28</sup> However, there is far less understanding of how to modify polymers to bind to functional proteins. Cells themselves consist of ~20–35% cytosol-stabilized protein by mass, depending on cell type.<sup>29</sup> In this crowded environment, it remains unclear how polymeric components of nanoparticles engage with proteins after endosomal escape. A better understanding of how proteins are compartmentalized by polymers into phase-separated domains could result in safer nanocarriers for protein based therapeutic drugs.

Tamasi *et al.* recently showed a unique approach to screen such polymer–protein hybrids (PPHs) using a learn–design–build–test paradigm for three model enzymes.<sup>30</sup> In this report, the authors prepared a series of heteropolymers that varied the (1) number of methacrylate monomer combinations (Fig. 4(A)), (2) balance of ionic, hydrophilic, and hydrophobic moieties (composition limited to  $\leq 70$  mol% hydrophobic and  $\leq 50$  mol% ionic monomer), and (3) targeted degree of polymerization (DP; from 50 to 200). PPHs were formed with horseradish peroxidase (HRP), glucose oxidase (GOx), and lipase (Lip) by thermal stress. The output objective was to evaluate retained enzyme activity (REA), defined as the ratio of activity level following thermal stress to its initial activity level. 500+

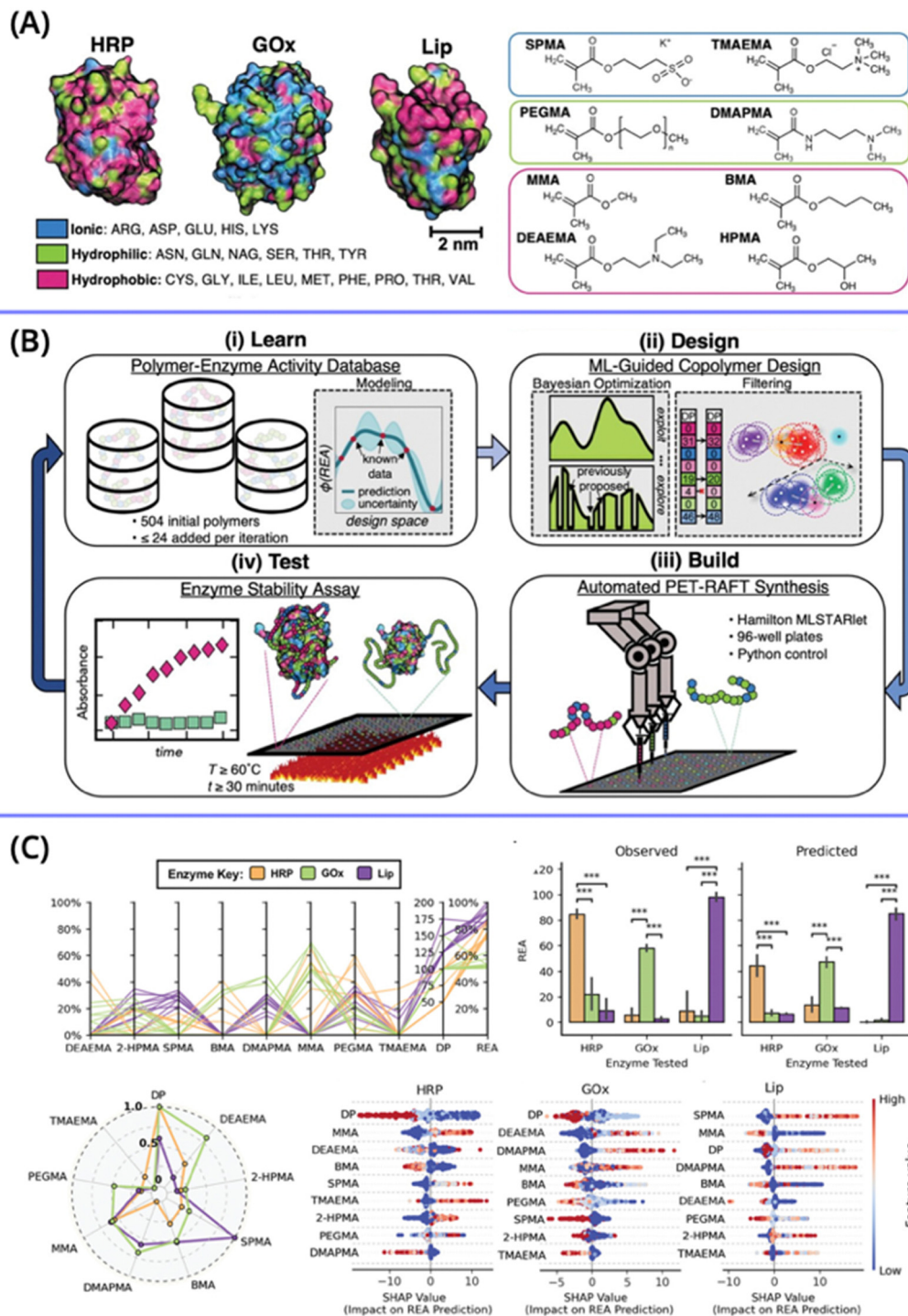
unique heteropolymers were prepared for enzymatic activity screening.

Closed-loop optimization was carried out by first training Gaussian process regression (GPR) models with a dataset of 504 initial polymers, followed by Bayesian optimization of the GPR model to down-select and identify lead polymer candidates for further synthesis campaigns. Evaluation of enzyme stability assays expanded the polymer–enzyme activity database for further model training and materials design (Fig. 4(B)). This workflow allowed the authors to better understand how chemical features influenced PPH performance of each protein of lead PPHs. While calculated SHAP values of REA showed expected trends, some unexpected relationships were revealed. For instance, smaller chain lengths and the hydrophobic monomer MMA were favorable for HRP, but the introduction of different hydrophobic monomers such as BMA was not beneficial (Fig. 4(C)). The authors proposed a possible mechanism of HRP stabilization as a chaperone-like assistance from shorter copolymer sequences that prevented structural refolding. SHAP analysis of GOx and Lip show distinct differences in heteropolymer design that were further improved by round-by-round Bayesian optimization coupled with experiments. This platform illustrated how ML workflows coupled with high-throughput materials experimentation can result in greater insight and speed to construct designer PPHs.

### 3. Outlook for pharmaceutical applications

In the sections above, we highlighted three use-case examples of how new small molecule and genetic drugs can be combined with customized polymers with guidance from ML tools. New datasets were collected from chemical and biological experiments to train ML models. In these instances, two central motifs emerged: (1) polymer chemistry techniques have advanced to enable exquisite control of virtually any excipient design, and (2) data-driven investigations can reveal non-intuitive attributes for polymer/payload stabilization and release.

However, translation of promising polymers/drug leads from the bench requires significant capital investment and resources on the path towards commercialization. In the remainder of this Highlight, we focus on nonviral gene therapy in particular, where there are numerous opportunities to produce more affordable and safer genetic medicines.<sup>31</sup> Some grand challenges stem from open-ended questions in molecular biology and nanomedicine, while others are more practical in terms of lab automation and establishing digital ecosystems for enabling material discovery. We discuss these topics and what may be needed for future pharmaceutical applications. More focused reviews on AI and nanomedicine,<sup>32</sup> nucleic acid therapeutics with polymer complexes,<sup>33</sup> and automation and data-driven design of polymer therapeutics<sup>34</sup> are available elsewhere. Furthermore, other biotherapeutics, such as ribonucleoproteins<sup>35</sup> or therapeutic peptides,<sup>5</sup> are outside the scope of this article.



**Fig. 4** Active learning enables rational design of polymer-protein hybrids (PPHs), comprising random copolymers with compositions that compatibilize protein surfaces. (A) Rendered surface chemistries of horseradish peroxidase (HRP), glucose oxidase (GOx), and lipase (Lip) whose amino acid attributes (ionic = blue, hydrophilic = green, hydrophobic = magenta) correspond to selected methacrylate chemistries. (B) Schematic of a "Learn-Design-Build-Test" PPH discovery paradigm, which includes Gaussian process regression surrogate models, Bayesian optimization, automated synthesis by a robotic platform, and high-throughput characterization assays. (C) Representative analysis reveals distinct priorities in copolymer features for each protein by normalized mean absolute SHAP explanations. Adapted with permission from ref. 30. Copyright 2022 Wiley John & Sons.

### 3.1. Diversity in cargo for nucleic acid therapies

There are several types of nucleic acid therapies, which vary in both therapeutic potential and delivery requirements. Table 1

summarizes the most prevalent classes based on technology advancement and usage in polymeric nanoparticle delivery. We emphasize notable chemical and biophysical considerations

**Table 1** Therapeutic nucleic acids considerations for polymer-driven nanoparticle delivery

Cargo	Delivery destination	Notable attributes	Select therapeutic product example(s) <sup>44</sup>
Plasmid DNA (pDNA)	Nucleus	<ul style="list-style-type: none"> <li>• Long (1000s bp<sup>a</sup>) double-stranded, circular molecule</li> <li>• Versatile, robust with low production cost from bacteria culture</li> <li>• Requires entry to restrictive nuclear barrier</li> </ul>	<ul style="list-style-type: none"> <li>• N/A; some DNA vaccines have been FDA approved for veterinary use such treating canine melanoma in 2010</li> </ul>
Antisense oligonucleotide (ASO)	Cytoplasm (RNA)	<ul style="list-style-type: none"> <li>• Short (~20 bases) single-stranded, linear molecule</li> <li>• Forms duplexes with RNA targets for promoting RNase degradation or for sterically blocking translation</li> <li>• Limited often by efficient internalization and endosomal escape</li> <li>• Chemical modifications are commonly used in ASO design</li> </ul>	<ul style="list-style-type: none"> <li>• Kynamro (FDA approved 2013)</li> <li>• Waylivra, Volanesoren (FDA approved 2019)</li> </ul>
Messenger RNA (mRNA)	Cytoplasm	<ul style="list-style-type: none"> <li>• Long (100–1000s bases) single-stranded, linear molecule</li> <li>• High expression, versatility, and therapeutic efficacy</li> <li>• Susceptible to RNase degradation, endosomal entrapment, and immune stimulation/response</li> <li>• Delivery vehicles are multicomponent, as general examples: lipid nanoparticles (PEGylated lipids, ionizable lipids, helper lipids, cholesterol), lipid/polymer hybrid nanoparticles (PEGylated lipids, cationic lipids, helper polymers), polyplexes (cationic polymers)</li> </ul>	<ul style="list-style-type: none"> <li>• Comirnaty, tozinameran (lipid nanoparticle-RNA FDA approved 2020)</li> <li>• mRNA-1273 (lipid nanoparticle-RNA FDA approved 2020)</li> </ul>
Small interfering RNA (siRNA)	Cytoplasm (RNA)	<ul style="list-style-type: none"> <li>• Short (15–30 bp) double-stranded, linear molecule</li> <li>• Forms complexes to bind and cut mRNA for blocking translation</li> <li>• Complementary and regulates the expression of a single target RNA to down-regulate protein expression levels</li> <li>• Many siRNA approaches motivated by cancer therapy</li> </ul>	<ul style="list-style-type: none"> <li>• Onpattro, patisiran (lipid nanoparticle-RNA FDA approved 2018)</li> <li>• Givlaari, Givosiran (GalNAc-siRNA conjugate FDA approved 2019)</li> <li>• Oxlumo, lumasiran (GalNAc-siRNA conjugate FDA approved 2020)</li> </ul>

<sup>a</sup> bp = base pairs, complementary repeat units in a nucleic acid molecule.

for featurization in ML models. Structural differences across nucleic acids are diverse and often overlooked when biological modifications are made for therapeutic function, but these changes may demand substantially different nanoparticles with features that are difficult to anticipate without large empirical datasets. Consequentially, one size does not fit all from a polymer design perspective, and re-examination of the structure and chemistry of DNA and RNA payloads may benefit the framework of molecular design and prediction capabilities with ML.

**3.1.1 DNA payloads.** DNA payloads are commonly used in gene therapy and offer the advantages of being relatively stable and capable of generating sizable quantities. DNA gene therapy cargoes are typically formulated as plasmids (pDNA), circular double-stranded DNA molecules. pDNA are easy and cheap to generate and can encode relatively large payloads. However, the large size (*i.e.*, molecular weight) of these pDNA payloads – often several kilobases long – can be incompatible with the limited packaging capacity (~4.5 kb) of viral delivery vehicles. With the addition of polycations in water, pDNA polyplexes can assemble into a variety of morphologies from doughnut-shaped to rod-like solids, depending on base pair (bp) length and associative interactions.<sup>36</sup> Furthermore, for pDNA to be transcribed, it must enter the nucleus; thus, a delivery vehicle needs to be capable of achieving nuclear import.

Oligonucleotides (typically defined as less than 100 bases) are short, single-stranded linear nucleic acids. In comparison to pDNA, they exhibit different complexation behavior that has been linked to differences in charge density, chain flexibility,

hydrophilicity, and helicity.<sup>37</sup> Antisense oligonucleotides (ASO) are a particular type of payload that is functional for gene silencing. ASO are short (~20 bases) and designed to bind to an endogenously expressed messenger RNA (mRNA) molecule. The ASO-mRNA duplex is then recognized and degraded by RNase H, resulting in reduced expression of the gene encoded by that mRNA.<sup>38</sup> Unlike pDNA, ASO do not need to enter the nucleus to have an effect: they silence gene expression through mRNA binding in the cytoplasm. However, delivery vehicles are still important for ASO, as they can facilitate uptake and protect against degradation.<sup>38</sup>

**3.1.2 RNA payloads.** RNA payloads come in several forms: mRNA can introduce genes into cells, while small interfering RNA (siRNA) can inhibit the expression of native genes. mRNA payloads have several advantages over pDNA cargo. They are typically smaller in size, which enables generally smaller polyplexes to be prepared. Additionally, because they are translated in the cytoplasm, they do not require nuclear entry to be effective therapeutically. However, mRNA suffers from disadvantages compared to DNA, such as low *in vivo* stability and high immunogenicity. These drawbacks can be mitigated through chemical modifications to the ribonucleotides, including modifications to nucleobases, ribose groups and phosphate backbones.<sup>39,40</sup> For mRNA, stability can be further improved through modifications to the 5' cap, as well as by optimizing secondary structure *via* changes to 5' and 3' untranslated regions and coding sequences.<sup>41–43</sup> It is unclear how such mRNA modifications impact polyplex assembly and delivery.

## Highlight

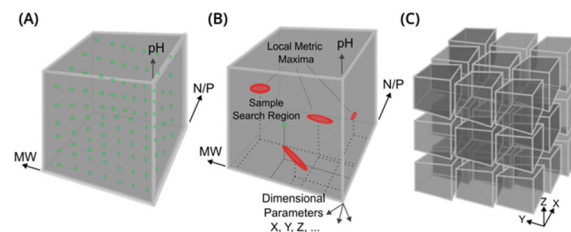
Gene silencing can be achieved through the delivery of short RNA molecules such as siRNA. siRNAs are short, double-stranded RNA molecules that can suppress gene expression through the action of a cytosolic protein complex known as an RNA-induced silencing complex (RISC). The siRNA sequence is designed to be complementary to part of an endogenous mRNA. The antisense strand of the siRNA guides the RISC to cleave the endogenous mRNA, thereby inhibiting protein production of the encoded gene. Like mRNA, siRNA nucleotides can be modified to protect against degradation. Delivery vehicles for siRNA can improve the cargo's stability and cellular uptake.<sup>38</sup> Hu *et al.* provide an extensive historic overview of therapeutic siRNA and a roadmap of their opportunities based on pre-clinical and clinical delivery platforms.<sup>45</sup>

### 3.2. Prospects for automation and cost reductions in gene therapy discovery and development

An emerging challenge and opportunity in MI is the “curse of dimensionality”:<sup>46</sup> the exponential increase in the size of a design space accompanied by the increase in the dimensionality of a problem.<sup>47</sup> As described earlier, gene therapy discovery and development with polymers is highly susceptible to this problem. Still, the rapid advances in lab automation hold immense potential for generating large and diverse datasets, rendering viable solutions and therefore reducing screening costs. There are numerous examples of the automation of polymer synthesis and combinatorial chemistry.<sup>48,49</sup> However, while this is certainly useful in reducing tedious, repetitive tasks, automation itself often requires significant effort to implement due to highly specialized tasks involving multi-step procedures.<sup>50</sup> Hence, we define “automation friendly” as being readily scalable processes with mechanically practical unit operations. We elaborate below on how one might address the immense chemical and biological parameter space and implement practical automated workflows.

**3.2.1 Chemical and biological parameter space.** In the case of polymer-driven gene delivery, optimization of chemistry and biology necessitates screening large number of samples with high degree of diversity. Exploring this variable design space requires synthesis and screening strategies to excel with a generalized approach. Experimental variability dealing with distinct classes of polymers, cargos, and biological targets create a high dimensional landscape for workflow pathways. The governing protocol needs to receive highly specific requirements for material synthesis and execute them in repeatable and technically sound way. Data quality and timeliness must be consistent so that structured chemical and biological data can be aggregated for training ML models. Increased performance over time is dependent on the metrics that are chosen by the trained predictive model.

Robotic systems can handle rote work and enable more complex operations to be completed in parallel. As an example, consider the use of liquid handling to prepare polyplexes. Input variables associated with mixing (*i.e.*, polymer concentration, cargo concentration, and solution salinity or pH) lead to enormous self-assembly outcomes that impact size, stability,



**Fig. 5** Visualization of how to discretize high dimensional spaces of forming polyplexes based polymer molecular weight (MW), N/P ratio, and solution pH as representative dimensional parameters. (A) Manually tested samples often only reveal behavior within narrow regions of sample space, so researchers may be left to interrogate overwhelming combinations of parameters. (B) To better probe system optima, a more tractable approach is to use robotic workflows to iteratively segment and parse hot spot regions as a basis for data collection. (C) The dimensionality is then deconstructed into more manageable spaces to enable further ML investigation, such as exploration (studying spaces where less is known but more can be learned) versus exploitation (studying spaces where more is known but can be highly improved in performance).

and nanoparticle dynamics. Fig. 5 highlights the challenges that occur with just three input parameters alone: polymer molecular weight, N/P (ratio of polymer/nucleic acid), and solution pH. A systematic screen of 10 molecular-weight polymer samples combined with a single cargo at 10 N/P ratios at 10 different pH levels translates to 1000 unique polyplexes that could in turn be conceivably plated in hours with automation (Fig. 5(A)). However, subdomains in the total dimensional parameter space can be more efficiently probed with prioritization assistance from MI techniques (Fig. 5(B)), identifying local optima (hot spots) of activity. Fig. 5(C) illustrates how these hot spots can be deconstructed further to allow the examination of additional dimensional factors. In this manner, nucleic acid stabilization and release can be better understood as a function of input parameters for subsequent workflows in polyplex characterization.

In the biological parameter space, automation and high-throughput screening has progressed significantly from its origins in small molecule drug discovery. Comprehensive reviews<sup>4,34,51</sup> present high-throughput evaluation of bioperformance in cellular and animal models. We direct readers to these works for more detailed perspectives. A common thread for these emerging techniques is establishing more autonomous workflows in the lab infrastructure (*e.g.*, plate preparation from libraries, assay standardization, or incorporation of non-invasive analytical techniques) and acceleration of decision making from the large quantity of collected data. We discuss these points further below.

**3.2.2 Automated workflow development.** Building a fully automated system of machines to tackle every step of a workflow remains a daunting task. A piecewise approach may be more realistic for smaller scale production. The determination of which procedures can best be automated and in what order is essential to streamline inefficiencies and bottlenecks. From the authors' experience, a design pyramid hierarchy for automated workflows can be useful in identifying specific pitfalls

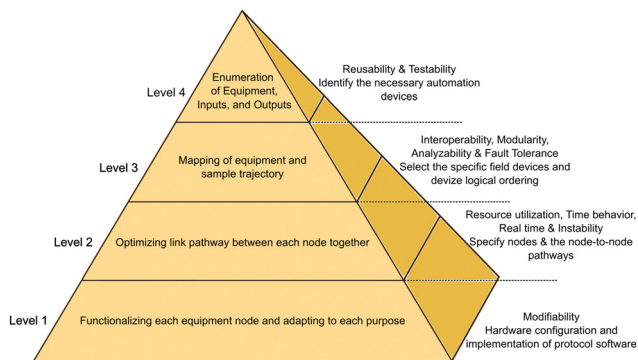


Fig. 6 Hierarchy of automated systems, organized by levels of practical and technical details for automation engineering. Levels 1 and 2 focus on modifiability of equipment nodes and optimally linking them together. Levels 3 and 4 involves interoperability and reusability to expand automation capabilities toward self-driving labs that can rapidly prototype materials discovery with minimal user involvement.

that may occur in scale up. Fig. 6 shows a hierarchy that illustrates areas of focus for building an automation system, with increasing levels of practical and technical details from top-to-bottom of the pyramid.

Following a top-down approach to meet a defined system's requirements, a map of equipment tasks should first be developed in a hierarchical manner. Each instrument should be chosen to perform in an automation friendly manner using (ideally) commercially available labware and consumables. Screening materials can be visualized at several levels, each of which have defined input and output (I/O) values. Building level 1 involves enumerating a list of lab-wide equipment. These I/O nodes should have a clearly defined purpose. Level 2 of the automation hierarchy necessitates organization of the nodes in a logical sequential manner. Process modularity, defined as inserting either redundant or alternative equipment at a process node, is crucial for engineering the pipeline to be adaptable and efficient. Creating a map of a work cell that depicts each step in the process with the I/O of each machine sub-system can be helpful.

Level 3 of the automation hierarchy focuses on optimizing resource utilization and the time course of each process. Delays or sample storage backlogs in the protocol are evidence of inefficiencies in the pipeline that can be resolved with redundancy removal or alternatives. Gantt charts may be useful in identifying timeline inefficiencies. From this, modifications to workflows can be gauged for their impact on sample production. Finally, at level 4, technical issues in the function of each process node are considered. Individual performance of each step is evaluated at the system level, whereas each device is detailed separately as its own finite system.

High-throughput synthesis and screening from automated system workflows must rely on an equally organized data management plan. Predictive ML models can find patterns in high dimensional data only if the data, and also its metadata can be connected. Metadata is data that contextualizes data by providing machine-readable descriptions and explainability.

We conclude below with our perspective on creating a digital infrastructure that can accommodate enormous chemical and biological datasets in polymer-based gene therapy.

### 3.3. Building digital infrastructure for machine actionable data

Although this article has emphasized various ML algorithms to predict small molecule and genetic drug delivery performance, MI workflows hinge on data that can be programmatically curated, stored, and used. Data pipelines allow data and metadata to be stored in a digital architecture. There are different strategies to generate data, ranging from monomer properties *in silico*<sup>51</sup> to polymer properties in experimentally prepared biomaterials.<sup>52</sup> Each requires different considerations for data acquisition, processing, and management to be machine actionable.<sup>53</sup> We discuss salient challenges for these three subjects below.

**3.3.1 Data acquisition.** The first step in establishing a data pipeline is data acquisition. This is the process of converting physical measurements into digital forms that can be accessed by computers or software. For polymer chemistry, recent examples include online size-exclusion chromatography,<sup>54</sup> inline NMR analysis,<sup>54–56</sup> UV-visible spectroscopy,<sup>57</sup> and real-time fluorescence tracking.<sup>58</sup> Here, we consider three data sources for any general measurement: assays that can be run without human intervention, assays that require human support, and external sources.

Ideally, a fully automated set-up is preferred when the run can be automatized with available hardware and an application programming interface (API). From the digital infrastructure point of view, data acquisition can often be integrated with software packages such as ChemOS,<sup>59</sup> HELAO,<sup>60</sup> or for physical simulations ChemOS 2.0.<sup>61</sup> These experiments offer high-throughput experimentation and high-throughput virtual screening capabilities with ML training.<sup>62</sup> This is a common theme for close-loop-optimization with ML examples in autonomous laboratories for inorganic,<sup>63</sup> organic<sup>64</sup> and polymer chemistry,<sup>19,65–67</sup> or even in nonviral delivery with lipid nanoparticles.<sup>68</sup>

In practice however, full automation often cannot be implemented economically. Experiments may have long turnaround times, reducing efficiencies in data acquisition. Hardware may lack a suitable API for I/O designation for development and optimization.<sup>69,70</sup> Standardized electronic laboratory notebooks (ELNs) can play an essential role. ELNs offer a user-friendly interface to record data through an API, so that data entry is both rapid and captured in a digitally useful form for MI applications.<sup>71</sup> Additional researcher input can be provided to improve experimental protocols, record quality control notes, and constrain data ingestion as appropriate.<sup>15</sup> Importantly, ELN adoption can allow access to negative results that are crucial for building balanced datasets. Negative data is usually not accessible or reported in public sources, a recognized problem in the field.<sup>72,73</sup> Commercially available software now offer APIs in combination with in-house development tools.<sup>74</sup>

The third source of data is published information. There are growing public datasets for polymers,<sup>75,76</sup> but much remains to be done compared to other materials genome or protein databases. Most published information is still not immediately machine actionable and requires substantial data extraction. Although large datasets of small molecules are widely available, these are often inadequate to extrapolate to polymers as they do not consider polymer synthesizability or other chemical constraints. Nevertheless, several polymer datasets exist. PoLy-Info<sup>77</sup> and Polymer Genome<sup>78</sup> are proprietary databases of existing polymers with focus in physical, mechanical, electrical, and chemical properties of, mostly, homopolymers. Querying these databases at scale is restricted and only a fraction of polymers and measured parameters are relevant to delivery applications. More varied datasets have been proposed by constructing virtual polymer using generative deep learning models, including PI1M,<sup>79</sup> polyBERT,<sup>80</sup> and the Open Macromolecular Genome (OMG).<sup>81</sup> In our opinion, the most relevant strategies to leverage these datasets are: (i) limited screening of restricted databases such as PoLyInfo and Polymer Genome based on known delivery vehicles, (ii) virtual screening of large virtual databases such as PI1M, OMG, and polyBERT (or fine-tuning their open-source generative models and representations), and (iii) for a polymer system with drug delivery potential, construct a dataset and/or ML model based on screening small molecule databases with polymer synthesizability constraints with an approach similar to the OMG.

With new AI technologies, datasets can be automatically extracted from text and figures, even from complex structures such as metal organics frameworks,<sup>82</sup> catalysts,<sup>83</sup> and chemical reaction schemes.<sup>84–86</sup> The rapid rise of generative models can also be used to aggregate molecule data from public resources.<sup>87</sup> Although not yet widely used in the field, deep learning has significant potential to accelerate polymer design in drug delivery. Beyond the discussed applications of generative models to define an accessible space for chemical design and suggest promising vehicle candidates, deep learning also has great potential for powerful polymer representation and generalization. Natural language architectures such as polyBERT learn useful representations for polymer property prediction<sup>80</sup> and generate linear random polymers. For more complex polymer architectures, graph representations<sup>88</sup> and extensions of BigSMILES<sup>89</sup> have been proposed which can be used as input to transformer or graph neural network architectures.

**3.3.2 Data processing.** Experimental data are routinely inhomogeneous. While much analytical data can be expressed in tabular form, increasingly, data streams involve complex images and image processing. In this regard, digitizing data is critical to effectively analyze results. For instance, there are reported data pipelines that incorporate image processing algorithms for expediting experiments.<sup>90</sup> From the software infrastructure perspective, it is critical to integrate a user-friendly interface that allows researchers to easily visualize and display processed data. Another critical need of establishing a digital infrastructure is computational power. At present, there are two main data and computational limitations for

polymer design for drug delivery. First, as discussed, the experimental data to build foundation models for polymer design is limited, restricting both the potential and computational requirements of large models compared to fields such as proteomics. Second, recent mechanistic modelling of polymers by molecular dynamics (MD) simulations<sup>91–93</sup> has progressed significantly; however, the molecular size and complexity of polymers and cargo limit the use of MD to a few polymers and simplified settings. As larger datasets are created alongside faster and more accurate MD proxies,<sup>92</sup> we expect these limitations to be mitigated. Future polymer design will have higher compute requirements for both mechanistic modelling and prediction of drug delivery properties.

A vital piece of the data processing pipeline includes meta-data collection. Metadata arises from different unit operations and parameters needed to understand results. Standardized formatting of data and metadata have been agreed upon by some materials communities, such as crystallography schema.<sup>94</sup> However, there is no consensus yet for nonviral drug delivery materials, including polymers. Conversations are yet needed to define and standardize metadata within the same lab group or organization to ensure data quality, consistency, and completeness.<sup>95</sup>

**3.3.3 Data management.** Data governance is a critical consideration to ensure short- and long-term accessibility for researchers and ML models alike.<sup>15,96</sup> A useful set of design directives is described by the FAIR (findable, accessible, interoperable, and reusable) data principles; FAIR goals establish guidelines for data pipelines to ensure the validity and readiness to use.<sup>97,98</sup> Scientific research data from federally funded grants increasingly need to adhere to FAIR data principles. The International Union of Pure and Applied Chemistry (IUPAC) has also begun to establish FAIR Chemistry guidelines for helping chemists manage their data.<sup>99</sup> Hardy and Heyse have discussed how FAIR data policies can benefit biotech startups.<sup>100</sup> In general, adoption of FAIR principles can over time lead to efficient retrospective analysis of legacy data, avoiding expensive manual interpretation of literature and archived data.

However, it remains unclear how to best address aspects of these principles while protecting intellectual property (IP), especially within companies where data is strategically valuable. In a thoughtful perspective,<sup>101</sup> Delannoy describes this problem between academic and industrial stakeholders: “the productivity of a project is measured by its capacity to transpose the research into products or services that will create business opportunities for the company. Patents are protected and even confidential for some time before being publicly published. For the academic partner, research is mostly evaluated by the scientific publications and communications that are published during a project. Papers are clearly public. It is therefore key to ensure a good balance between the protection of IP that the company needs and the objective to publish that the university seeks.” Biopharma groups have considered aspects of this as part of their digital transformation process,<sup>102</sup> but it still remains unclear how to best handle this issue moving forward.

## 4. Conclusions

We have provided an outlook on nonviral delivery using polymer-based therapeutic nanocarriers for small molecule and genetic drugs. In the rapidly growing field of gene therapy, the remarkable progress of integrating polymer chemistry with ML offers a glimpse of the next evolution in design rules and practice for chemists. Such data-driven approaches can lay the groundwork for advancing more genetic treatments to market at a rate that was inconceivable decades ago.

Many approved gene therapies to date can be traced to academic laboratories or small companies.<sup>103</sup> While important advances have been made, scaling of these efforts has been hampered by both the complexity of the data and the difficulty of integrating and interpreting disparate data streams. MI methods offer enormous potential because of the possibility of scaling experimental and computational methods, including polymer-cargo formulation. In our own experience, we believe that MI can be leveraged for the efficient design of polymer nanoparticles for nonviral gene therapy. Our SAYER<sup>TM</sup> platform<sup>104</sup> connects high-throughput polymer chemistry, polyplex nanoparticle characterization, screening/delivery *in vitro* and *in vivo*, and predictive polymer design guided by ML and AI. Fit-for-purpose delivery vehicles can be rapidly produced, down-selected for cargo and tissue specificity, and assayed in a time- and cost-effective manner.

Advances in the integration of experimental and MI methods will continue to evolve over the coming decades. The existing R&D landscape will likely develop in profound and unexpected ways. The rise of generative AI is one such high-impact example. We have attempted above to demonstrate how this integration of experimental and MI methods can be used to discover unexpected correlations and new insights into structure–property relationships. For the field of gene therapy in particular, involving highly complex interplay between physical, biological, and clinical factors, such modalities can fuel more rapid innovation and affordable solutions that benefit people worldwide.

## Glossary of MI terms

- **LGBM:** Light gradient boosting machine, a tree-based machine learning algorithm.
- **SHAP:** SHapley Additive exPlanations, a method used to understand the role of the features in an ML method.
- **PCA:** Principal component analysis, a method to transform high-dimensional data to low-dimensions.
- **tSNE:** T-distributed Stochastic Neighbor Embedding, a clustering algorithm used to visualize high-dimensional data in 2 or 3 dimensions.
- **BO:** Bayesian optimization, an algorithm typically used to optimize properties expensive or difficult to evaluate.
- **GPR:** Gaussian process regression.
- **API:** Application programming interface, a protocol that allows to read, write and/or manage different instances programmatically, such as software applications or a piece of hardware.

- **ELN:** Electronic laboratory notebooks, a software application that allows experimentalists to keep a digital logbook.

## Author contributions

Each section above was written by the following author(s) in parentheses: Introduction (J. M. T.), Small Molecule Drugs (T. T.), Nucleic Acid and Ribonucleoprotein Delivery (S. R. P., J. V. R.), Polymer–Protein Hybrids (J. M. T.), Diversity in Cargo for Nucleic Acid Therapies (U. A. A., J. D. F., N. J. S.), Prospects for Automation and Cost Reductions in Gene Therapy Discovery and Development (M. S.), Building Digital Infrastructure for Machine Actionable Data (T. T., F. O.), Conclusions (J. M. T.). All authors participated equally in the team discussion, editing, and proofreading process. J. M. T. is responsible for the conceptualization and final revisions of the article.

## Conflicts of interest

The authors declare the following competing financial interest(s): all authors have an equity interest in Nanite, Inc.

## Acknowledgements

We acknowledge Shashi Murthy, Thomas Neenan, and Sean Kevlahan for helpful discussions, manuscript feedback, and support in preparation of this article.

## Notes and references

- 1 National Science Foundation Workshop: Frontiers in Polymer Science and Engineering, Aug. 17, 2017, <https://sites.google.com/a/umn.edu/nsf-polymer-workshop/report> (accessed Oct. 30, 2023).
- 2 T. Lodge, *Phys. Today*, 2017, **70**, 10–12.
- 3 K. Park, *J. Controlled Release*, 2014, **190**, 3–8.
- 4 R. Kumar, C. F. Santa Chalarca, M. R. Bockman, C. V. Bruggen, C. J. Grimme, R. J. Dalal, M. G. Hanson, J. K. Hexum and T. M. Reineke, *Chem. Rev.*, 2021, **121**, 11527–11652.
- 5 H. Acar, J. M. Ting, S. Srivastava, J. L. LaBelle and M. V. Tirrell, *Chem. Soc. Rev.*, 2017, **46**, 6553–6569.
- 6 T. Bus, A. Traeger and U. S. Schubert, *J. Mater. Chem. B*, 2018, **6**, 6904–6918.
- 7 R. B. Grubbs and R. H. Grubbs, *Macromolecules*, 2017, **50**, 6979–6997.
- 8 E. Blasco, M. B. Sims, A. S. Goldmann, B. S. Sumerlin and C. Barner-Kowollik, *Macromolecules*, 2017, **50**, 5215–5252.
- 9 C. Chen, D. Y. W. Ng and T. Weil, *Prog. Polym. Sci.*, 2020, **105**, 101241.
- 10 C. C. M. Sproncken, J. R. Magana and I. K. Voets, *ACS Macro Lett.*, 2021, **10**, 167–179.
- 11 M. Reis, F. Gusev, N. G. Taylor, S. H. Chung, M. D. Verber, Y. Z. Lee, O. Isayev and F. A. Leibfarth, *J. Am. Chem. Soc.*, 2021, **143**, 17677–17689.
- 12 J. M. Ting, S. Tale, A. A. Purchel, S. D. Jones, L. Widanapathirana, Z. P. Tolstyka, L. Guo, S. J. Guillaudeu, F. S. Bates and T. M. Reineke, *ACS Cent. Sci.*, 2016, **2**, 748–755.
- 13 J. L. Mann, C. L. Maikawa, A. A. Smith, A. K. Grosskopf, S. W. Baker, G. A. Roth, C. M. Meis, E. C. Gale, C. S. Liong, S. Correa, D. Chan, L. M. Stapleton, A. C. Yu, B. Muir, S. Howard, A. Postma and E. A. Appel, *Sci. Transl. Med.*, 2020, **12**, eaba6676.
- 14 Council, “Data Challenges Are Halting AI Projects, IBM Executive Says” Wall Street Journal, 2019, <https://www.wsj.com/articles/data-challenges-are-halting-ai-projects-ibm-executive-says-11559035800> (accessed: Oct. 30, 2023).

- 15 C. Willoughby and J. G. Frey, *Digital Discovery*, 2022, **1**, 183–194.
- 16 J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, J. Analytis, I. Dabo, D. M. DeLongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton and J.-C. Zhao, *npj Comput. Mater.*, 2019, **5**, 41.
- 17 D. J. Audus and J. J. de Pablo, *ACS Macro Lett.*, 2017, **6**, 1078–1082.
- 18 A. J. Gormley and M. A. Webb, *Nat. Rev. Mater.*, 2021, **6**, 642–644.
- 19 T. B. Martin and D. J. Audus, *ACS Polym. Au*, 2023, **3**, 239–258.
- 20 M. N. O'Brien, W. Jiang, Y. Wang and D. M. Loffredo, *J. Controlled Release*, 2021, **336**, 144–158.
- 21 O. M. H. Salo-Ahen, I. Alanko, R. Bhadane, A. M. J. J. Bonvin, R. V. Honorato, S. Hossain, A. H. Juffer, A. Kabedev, M. Lahtela-Kakkonen, A. S. Larsen, E. Lescrinier, P. Marimuthu, M. U. Mirza, G. Mustafa, A. Nunes-Alves, T. Pantisar, A. Saadabadi, K. Singaravelu and M. Vanmeert, *Processes*, 2020, **9**, 71.
- 22 P. Bannigan, Z. Bao, R. J. Hickman, M. Aldeghi, F. Häse, A. Aspuru-Guzik and C. Allen, *Nat. Commun.*, 2023, **14**, 35.
- 23 H. Gasmi, F. Siepmann, M. C. Hamoudi, F. Danede, J. Verin, J.-F. Willart and J. Siepmann, *Int. J. Pharm.*, 2016, **514**, 189–199.
- 24 M. Tirrell, *AIChE J.*, 2005, **51**, 2386–2390.
- 25 R. Kumar, N. Le, F. Oviedo, M. E. Brown and T. M. Reineke, *JACS Au*, 2022, **2**, 428–442.
- 26 N. D. Posey, C. R. Hango, L. M. Minter and G. N. Tew, *Bioconjugate Chem.*, 2018, **29**, 2679–2690.
- 27 G. Chen, A. A. Abdeen, Y. Wang, P. K. Shahi, S. Robertson, R. Xie, M. Suzuki, B. R. Pattnaik, K. Saha and S. Gong, *Nat. Nanotechnol.*, 2019, **14**, 974–980.
- 28 Y.-W. Lee, D. C. Luther, R. Goswami, T. Jeon, V. Clark, J. Elia, S. Gopalakrishnan and V. M. Rotello, *J. Am. Chem. Soc.*, 2020, **142**, 4349–4355.
- 29 A. B. Fulton, *Cell*, 1982, **30**, 345–347.
- 30 M. J. Tamasi, R. A. Patel, C. H. Borca, S. Kosuri, H. Mugnier, R. Upadhy, N. S. Murthy, M. A. Webb and A. J. Gormley, *Adv. Mater.*, 2022, **34**, 2201809.
- 31 C. Sheridan, *Nat. Biotechnol.*, 2023, **41**, 737–739.
- 32 N. Serov and V. Vinogradov, *Adv. Drug Delivery Rev.*, 2022, **184**, 114194.
- 33 U. Lächelt and E. Wagner, *Chem. Rev.*, 2015, **115**, 11043–11078.
- 34 R. Upadhy, S. Kosuri, M. Tamasi, T. A. Meyer, S. Atta, M. A. Webb and A. J. Gormley, *Adv. Drug Delivery Rev.*, 2021, **171**, 1–28.
- 35 S. Zhang, J. Shen, D. Li and Y. Cheng, *Theranostics*, 2021, **11**, 614–648.
- 36 U. K. Laemmli, *Proc. Natl. Acad. Sci. U. S. A.*, 1975, **72**, 4288–4292.
- 37 J. R. Viereg, M. Lueckheide, A. B. Marciel, L. Leon, A. J. Bologna, J. R. Rivera and M. V. Tirrell, *J. Am. Chem. Soc.*, 2018, **140**, 1632–1638.
- 38 J. C. Kaczmarek, P. S. Kowalski and D. G. Anderson, *Genome Med.*, 2017, **9**, 60.
- 39 K. Karikó, H. Muramatsu, F. A. Welsh, J. Ludwig, H. Kato, S. Akira and D. Weissman, *Mol. Ther.*, 2008, **16**, 1833–1840.
- 40 M. S. D. Kormann, G. Hasenpusch, M. K. Aneja, G. Nica, A. W. Flemmer, S. Herber-Jonat, M. Huppmann, L. E. Mays, M. Illenyi, A. Schams, M. Gries, I. Bittmann, R. Handgretinger, D. Hartl, J. Rosenecker and C. Rudolph, *Nat. Biotechnol.*, 2011, **29**, 154–157.
- 41 M. Strenkowska, R. Grzela, M. Majewski, K. Wnek, J. Kowalska, M. Lukaszewicz, J. Zuberek, E. Darzynkiewicz, A. N. Kuhn, U. Sahin and J. Jemielity, *Nucleic Acids Res.*, 2016, **44**, 9578–9590.
- 42 K. Leppke, G. W. Byeon, W. Kladwang, H. K. Wayment-Steele, C. H. Kerr, A. F. Xu, D. S. Kim, V. V. Topkar, C. Choe, D. Rothschild, G. C. Tiu, R. Wellington-Oguri, K. Fujii, E. Sharma, A. M. Watkins, J. J. Nicol, J. Romano, B. Tunguz, F. Diaz, H. Cai, P. Guo, J. Wu, F. Meng, S. Shi, E. Participants, P. R. Dormitzer, A. Solórzano, M. Barna and R. Das, *Nat. Commun.*, 2022, **13**, 1536.
- 43 D. M. Mauger, B. J. Cabral, V. Presnyak, S. V. Su, D. W. Reid, B. Goodman, K. Link, N. Khatwani, J. Reyniers, M. J. Moore and I. J. McFadyen, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 24075–24083.
- 44 J. A. Kulkarni, D. Witzigmann, S. B. Thomson, S. Chen, B. R. Leavitt, P. R. Cullis and R. Van Der Meel, *Nat. Nanotechnol.*, 2021, **16**, 630–643.
- 45 B. Hu, L. Zhong, Y. Weng, L. Peng, Y. Huang, Y. Zhao and X.-J. Liang, *Signal Transduction Targeted Ther.*, 2020, **5**, 101.
- 46 R. Bellman, *Science*, 1966, **153**, 34–37.
- 47 A. A. Volk, R. W. Epps, D. T. Yonemoto, B. S. Masters, F. N. Castellano, K. G. Reyes and M. Abolhasani, *Nat. Commun.*, 2023, **14**, 1403.
- 48 S. Oliver, L. Zhao, A. J. Gormley, R. Chapman and C. Boyer, *Macromolecules*, 2019, **52**, 3–23.
- 49 M. Tamasi, S. Kosuri, J. DiStefano, R. Chapman and A. J. Gormley, *Adv. Intell. Syst.*, 2020, **2**, 1900126.
- 50 M. Christensen, L. P. E. Yunker, P. Shiri, T. Zepel, P. L. Prieto, S. Grunert, F. Bork and J. E. Hein, *Chem. Sci.*, 2021, **12**, 15473–15490.
- 51 L. Schneider, M. Schwarting, J. Mysona, H. Liang, M. Han, P. M. Rauscher, J. M. Ting, S. Venkatram, R. B. Ross, K. J. Schmidt, B. Blaiszik, I. Foster and J. J. de Pablo, *Mol. Syst. Des. Eng.*, 2022, **7**, 1611–1621.
- 52 S. M. McDonald, E. K. Augustine, Q. Lanners, C. Rudin, L. Catherine Brinson and M. L. Becker, *Nat. Commun.*, 2023, **14**, 1–11.
- 53 J. Kimmig, S. Zechel and U. S. Schubert, *Adv. Mater.*, 2021, **33**, 2004940.
- 54 J. Van Herck, I. Abeysekera, A.-L. Buckinx, K. Cai, J. Hooker, K. Thakur, E. Van De Reydt, P.-J. Voort, D. Wyers and T. Junkers, *Digital Discovery*, 2022, **1**, 519–526.
- 55 M. Rubens, J. Van Herck and T. Junkers, *ACS Macro Lett.*, 2019, **8**, 1437–1441.
- 56 S. T. Knox, S. Parkinson, R. Stone and N. J. Warren, *Polym. Chem.*, 2019, **10**, 4774–4778.
- 57 F. Lauterbach and V. Abetz, *Macromol. Rapid Commun.*, 2020, **41**, 2000029.
- 58 J. Lee, P. Mulay, M. J. Tamasi, J. Yeow, M. M. Stevens and A. J. Gormley, *Digital Discovery*, 2023, **2**, 219–233.
- 59 L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein and A. Aspuru-Guzik, *Sci. Robot.*, 2018, **3**, eaat5559.
- 60 F. Rahmanian, J. Flowers, D. Guevarra, M. Richter, M. Fichtner, P. Donnelly, J. M. Gregoire and H. S. Stein, *Adv. Mater. Interfaces*, 2022, **9**, 2101987.
- 61 M. Sim, M. Ghazi Vakili, F. Strieth-Kalthoff, H. Hao, R. Hickman, S. Miret, S. Pablo-Garcia and A. Aspuru-Guzik, *ChemRxiv*, 2023, DOI: [10.26434/chemrxiv-2023-v2khf](https://doi.org/10.26434/chemrxiv-2023-v2khf).
- 62 J. K. Wilson, Doctor of Philosophy, Drexel University, 2022.
- 63 C. B. Wahl, M. Aykol, J. H. Swisher, J. H. Montoya, S. K. Suram and C. A. Mirkin, *Sci. Adv.*, 2021, **7**, eabj5505.
- 64 N. H. Angello, V. Rathore, W. Beker, A. Wołos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski and M. D. Burke, *Science*, 2022, **378**, 399–405.
- 65 A. Vriza, H. Chan and J. Xu, *Chem. Mater.*, 2023, **35**, 3046–3056.
- 66 O. Queen, G. A. McCarver, S. Thatigotla, B. P. Abolins, C. L. Brown, V. Maroulas and K. D. Vogiatzis, *npj Comput. Mater.*, 2023, **9**, 90.
- 67 S. T. Knox, S. J. Parkinson, C. Y. P. Wilding, R. A. Bourne and N. J. Warren, *Polym. Chem.*, 2022, **13**, 1576–1585.
- 68 R. J. Hickman, P. Bannigan, Z. Bao, A. Aspuru-Guzik and C. Allen, *Matter*, 2023, **6**, 1071–1081.
- 69 M. Seifrid, R. Pollice, A. Aguilar-Granda, Z. Morgan Chan, K. Hotta, C. T. Ser, J. Vestfrid, T. C. Wu and A. Aspuru-Guzik, *Acc. Chem. Res.*, 2022, **55**, 2454–2466.
- 70 M. Abolhasani and E. Kumacheva, *Nat. Synth.*, 2023, **2**, 483–492.
- 71 M. Alexopoulos and T. Tombe, *J. Monet. Econ.*, 2012, **59**, 269–285.
- 72 M. P. Maloney, C. W. Coley, S. Genheden, N. Carson, P. Helquist, P.-O. Norrby and O. Wiest, *Org. Lett.*, 2023, **25**, 2945–2947.
- 73 P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.
- 74 S. G. Higgins, A. A. Nogiwa-Valdez and M. M. Stevens, *Nat. Protoc.*, 2022, **17**, 179–189.
- 75 T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania and R. Ramprasad, *Sci. Data*, 2016, **3**, 160012.
- 76 D. J. Walsh, W. Zou, L. Schneider, R. Mello, M. E. Deagen, J. Mysona, T.-S. Lin, J. J. De Pablo, K. F. Jensen, D. J. Audus and B. D. Olsen, *ACS Cent. Sci.*, 2023, **9**, 330–338.
- 77 S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu and M. Yamazaki, in 2011 International Conference on Emerging Intelligent Data and Web Technologies, IEEE, Tirana, Albania, 2011, pp. 22–29.
- 78 C. Kim, A. Chandrasekaran, T. D. Huan, D. Das and R. Ramprasad, *J. Phys. Chem. C*, 2018, **122**, 17575–17585.

- 79 R. Ma and T. Luo, *J. Chem. Inf. Model.*, 2020, **60**, 4684–4690.
- 80 C. Kuenneth and R. Ramprasad, *Nat. Commun.*, 2023, **14**, 4099.
- 81 S. Kim, C. M. Schroeder and N. E. Jackson, *ACS Polym. Au*, 2023, **3**, 318–330.
- 82 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, *J. Am. Chem. Soc.*, 2023, **145**, 18048–18062.
- 83 H. Joshi, N. Wilde, T. S. Asche and D. Wolf, *Chem. Ing. Tech.*, 2022, **94**, 1645–1654.
- 84 R. B. Tchoua, K. Chard, D. Audus, J. Qin, J. De Pablo and I. Foster, *Procedia Comput. Sci.*, 2016, **80**, 386–397.
- 85 P. Shetty and R. Ramprasad, *iScience*, 2021, **24**, 101922.
- 86 P. Shetty, A. C. Rajan, C. Kuenneth, S. Gupta, L. P. Panchumarti, L. Holm, C. Zhang and R. Ramprasad, *npj Comput. Mater.*, 2023, **9**, 52.
- 87 D. M. Anstine and O. Isayev, *J. Am. Chem. Soc.*, 2023, **145**, 8736–8750.
- 88 M. Aldeghi and C. W. Coley, *Chem. Sci.*, 2022, **13**, 10486–10498.
- 89 L. Schneider, D. Walsh, B. Olsen and J. De Pablo, *ChemRxiv*, 2023, DOI: [10.26434/chemrxiv-2023-xv1kf](https://doi.org/10.26434/chemrxiv-2023-xv1kf).
- 90 J. Zhang, C. Li, M. M. Rahaman, Y. Yao, P. Ma, J. Zhang, X. Zhao, T. Jiang and M. Grzegorzczek, *Artif. Intell. Rev.*, 2022, **55**, 2875–2944.
- 91 M. A. Webb, N. E. Jackson, P. S. Gil and J. J. de Pablo, *Sci. Adv.*, 2020, **6**, eabc6216.
- 92 R. Feng, H. Tran, A. Toland, B. Chen, Q. Zhu, R. Ramprasad and C. Zhang, *arXiv*, 2023, DOI: [10.48550/arXiv.2309.00585](https://doi.org/10.48550/arXiv.2309.00585).
- 93 N. E. Jackson, B. M. Savoie, A. Statt and M. A. Webb, *J. Chem. Theory Comput.*, 2023, **19**, 4335–4337.
- 94 S. R. Hall and B. McMahon, *International Tables for Crystallography*, Springer Science & Business Media, 2005.
- 95 J. M. Ting and C. E. Lipscomb, *Pure Appl. Chem.*, 2022, **94**, 637–642.
- 96 B. G. Pelkie and L. D. Pozzo, *Digital Discovery*, 2023, **2**, 544–556.
- 97 M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. Da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. Van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. Van Der Lei, E. Van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, *Sci. Data*, 2016, **3**, 160018.
- 98 E. T. Inau, J. Sack, D. Waltemath and A. A. Zeleke, *JMIR Res. Protoc.*, 2021, **10**, e22505.
- 99 I. Bruno, S. Coles, W. Koch, L. McEwen, F. Meyers and S. Stall, *Chem. Int.*, 2021, **43**, 12–16.
- 100 K. Hardy and S. Heyse, *Nat. Biotechnol.*, 2023, **41**, 1060–1061.
- 101 J.-Y. P. Delannoy, *ACS Polym. Au*, 2022, **2**, 137–146.
- 102 J. Wise, A. G. De Barron, A. Splendiani, B. Balali-Mood, D. Vasant, E. Little, G. Mellino, I. Harrow, I. Smith, J. Taubert, K. Van Bochove, M. Romacker, P. Walgemoed, R. C. Jimenez, R. Winnenburg, T. Plasterer, V. Gupta and V. Hedley, *Drug Discovery Today*, 2019, **24**, 933–938.
- 103 K. N. Vokinger, J. Avorn and A. S. Kesselheim, *N. Engl. J. Med.*, 2023, **388**, 292–295.
- 104 Nanite, <https://www.nanitebio.com/> (accessed Oct. 30, 2023).