CellPress

Matter

Matter of Opinion

Machine learning for next-generation nanotechnology in healthcare

Andzelika Lorenc,[1,2,5] Bárbara B. Mendes,[3,4,5] João Conniot,[3,4,5] Diana P. Sousa,[3,4,5] João Conde,[3,4,*] and Tiago Rodrigues[1,*]

Nanotechnology for healthcare is coming of age, but automating the design of composite materials poses unique challenges. Although machine learning is supporting groundbreaking discoveries in materials science, new initiatives leveraging learned patterns are required to fully realize the promise of nanodelivery systems and accelerate development pipelines.

Nanotechnology has seen numerous translational applications over diverse economy sectors, but only recently has it taken healthcare by storm.[1] For example, the massively deployed Moderna and BioNTech/Pfizer COVID-19 vaccines use organic nanoparticles to deliver SARS-CoV-2 mRNA.[2] The pursuit of this novel immunization modality has positioned nanotechnology at the center of attention, providing ample clinical validation and further motivating its exploitation in disparate disease areas, e.g., cancer. In this regard, billions of dollars spent in basic/translational nanotechnology have, over the years, allowed a reasonable understanding of the design principles driving efficacy.[1] Still, much remains to be explored. The anticipated shift to nanotechnology-centered molecular medicine urges the need for an innovative suite of computational tools that effectively harness the growing amount of information in this space. A data-driven (r)evolution, similar to what we are currently witnessing in chemistry and biology, might not be too distant. We argue that predictive modeling and the *de novo* design of composite nanodelivery systems will become a reality and ultimately endorse a new era in nanotechnology research. Herein, we critically discuss how machine learning (ML) can reshape next-generation drug delivery and the three challenges that must be addressed to enable continuous innovation through discriminative/generative nanotechnology.

Challenge 1: Standardized data reporting

Quality data are unavoidably the centerpiece of any ML tool, and the current lack of standardized reporting practices in nanobiotechnology and nanomedicine is a known issue (Figure 1).[3,4] This hinders reproducibility and meaningful comparative studies, despite a recent community effort to regulate and improve transparency in the disclosed materials.[4] For example, the physicochemical properties (e.g., dimension, shape, surface charge, targeting agent density, and composition), administered dose, and loading in drug delivery systems are cornerstones to modulate pharmacokinetics and efficacy. However, their reporting heterogeneity or lack of explicit information in manuscripts jeopardizes the gain of momentum in nanomedical research.[4] Further, we argue the exact composition, injected volume, concentration, and route of administration need to be accurately reported but are only sparingly discriminated. Multiple studies also describe the amount of only one component in the delivery system (e.g., iron or encapsulated drug dosage). Thus, normalizing the others by body weight in *in vivo* studies is virtually impossible. Paralleling deficiencies in the characterization of nanodelivery materials, one equally finds shortcomings in the report of assay endpoints. Delivery efficacy and tumor accumulation are usually provided as a percentage of initial dose/tumor mass in gram (%ID/g). This is only useful if the initial dose and tumor mass are also reported, which rarely is the case. Further, the percentage of tumor reduction constitutes an endpoint normalization, yielding no tractable information on either delivery efficiency or actual volume change. Finally, one must bear in mind that animal models are a proxy of an actual disease setting, and experiments must be carefully designed to ensure meaningfulness. Otherwise, they may inappropriately reproduce the disease and tumor microenvironment, as in the use of xenograft/allograft heterotopic models instead of orthotopic ones (e.g., lung cells injected subcutaneously).

Overall, the design of nanomaterials is intricate, and the established practices in experimental characterization are manifestly insufficient to support translation in healthcare on a wider scale.[5] Perpetuating those practices will ultimately curb innovation and research

[1]Instituto de Investigação do Medicamento (iMed), Faculdade de Farmácia, Universidade de Lisboa, Av. Prof. Gama Pinto, 1649-003 Lisboa, Portugal

[2]Department of Biopharmacy, Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, Jurasza 2, 85-089 Bydgoszcz, Poland

[3]NOVA Medical School, Faculdade de Ciências Médicas, Universidade Nova de Lisboa, 1169-056 Lisboa, Portugal

[4]Centre for Toxicogenomics and Human Health, Genetics, Oncology and Human Toxicology, NOVA Medical School, Faculdade de Ciências Médicas, Universidade Nova de Lisboa, 1169 056 Lisboa, Portugal

[5]These authors contributed equally

*Correspondence: joao.conde@nms.unl.pt (J.C.), tiago.rodrigues@ff.ulisboa.pt (T.R.)

https://doi.org/10.1016/j.matt.2021.09.014

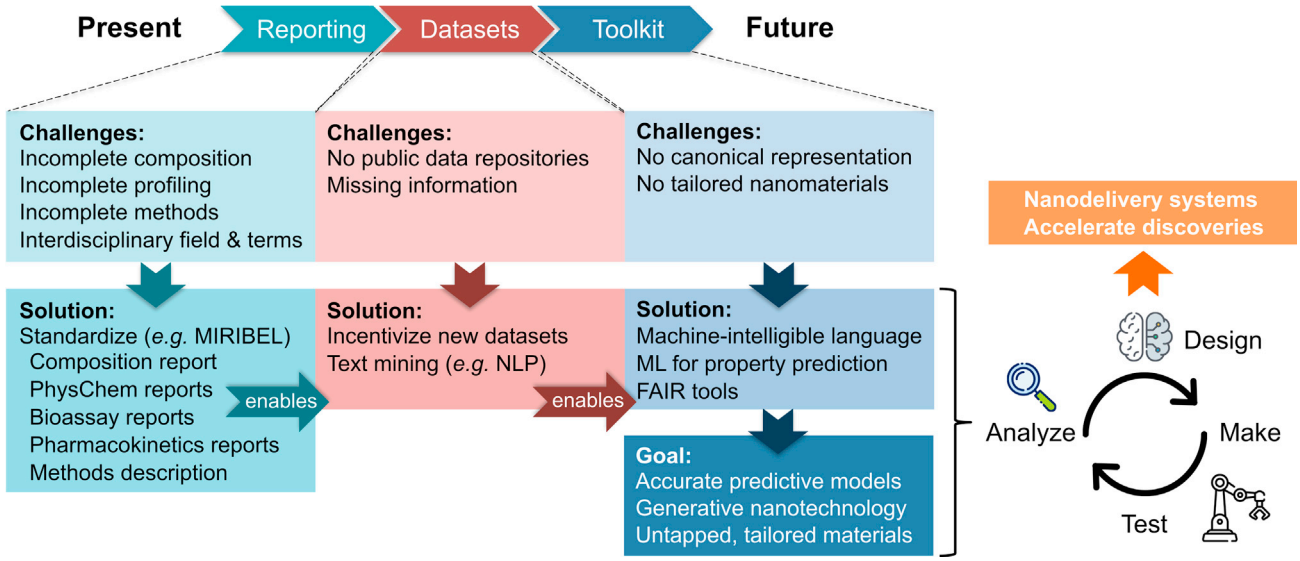

Figure 1. Schematics of the key challenges to address and the proposed solutions to enable continuous innovation and accelerate productivity in the development of nanodelivery systems

Standardization of reported data, as suggested in the MIRIBEL guidelines,[4] will enable the construction of high-quality datasets for statistical modeling. These include harmonizing the reporting of composition (e.g., loadings, coatings), physicochemical properties (e.g., size, zeta potential, shape), bioassay readouts, pharmacokinetics, and methods description. We envisage that the use of natural language processing (NLP) principles can help in constructing open source databases that can be used for myriad statistical analyses and support the development of new single-entity and/or composite materials. Also, the creation of a new language for canonical representation of composite drug delivery systems will allow developing ML models for property prediction and establish generative nanotechnology as a paradigm for the *in silico* design of nanomaterials. Said tools should follow the "findable, accessible, interoperable, and reusable" (FAIR) principles to enhance democratization and may find integration in robotic systems to close the make-design-test cycle.

throughput. We envisage that a shift to robust reporting standards will be instrumental for automating the development of nanodelivery materials.

Challenge 2: Complete datasets

Reporting completeness, as urged above, will then enable the construction of nanotechnology databases akin to ChEMBL (Figure 1).[6] To emphasize the critical nature of quality and complete data, we extracted composition, physicochemical, pharmacokinetics, and dosing information for iron oxide nanoparticles reported in 315 research manuscripts published between 2006 and 2019 in reputed nanotechnology journals. From those, 68% did not report the size, shape, or zeta potential of the nanoparticles. Further, only 1% and 31% presented a pharmacokinetics profile (with elimination/distribution half-lives and delivery efficiency) and dosing information (route and dose), respectively. The observed trends are apparently transferred to other nanoparticle types. In 322 gold and 257 silica nanoparticle research manuscripts, we found 51% and 45% of them missing full physicochemical characterization, respectively. Identical percentages of missing pharmacokinetics and dosing data were found. Together, this highlights a deep-rooted limitation in the nanotechnology field that must be addressed. Until then, inputting or discarding potentially valuable information will be required for modeling, which is far from ideal.

While a publicly available resource will likely remain inaccessible in the coming years—even with adequate reporting standards in place—we envisage that a continuous and concerted community effort will be key toward that end. Those efforts might be further assisted by natural language processing and deep learning techniques with the goal of accelerating the extraction of information from the scientific literature.[7] By encompassing multiple unexplored data patterns, those data resources are expected to endorse automated processes and support the implementation of ML tools that allow more efficient experiment prioritizations.

Challenge 3: A machine-readable nanotechnology language

While predictive modeling[8,9] can be executed with quality datasets and established heuristics, generative design of composite materials requires the development of new toolkits. In small-molecule discovery, the SMILES or SELFIES[10] languages encode atom connectivity, which implicitly hardwires all physicochemical and biological properties for a given entity. By learning this language, a computer is then able to programmatically generate new words/molecules (as strings of characters) according to a

probability distribution for each newly added character. In doing so, researchers become armed with a powerful means to virtually access new chemical matter and more efficiently explore a vast search space. We argue that a similar approach can be pursued to generate and tailor composite delivery materials (Figure 1). Considering that information on constituents, including which entities, their percentage, and/or their concentration, is key to determining all the underlying physicochemical and biological properties, it becomes essential to devise a new language and ontology for canonical representation of composite material systems—both already reported and imagined by a computer. Said language ought to holistically represent the nanomaterial and thus be transferable to any use case aside from the drug delivery systems we focus on here. Once this technology is in hand, the research community will gain access to an untapped concept for the *de novo* design of composite materials. If employed correctly, we expect that such ML models—which might be available in a short/medium term—could impact nanotechnology for healthcare similarly to how they are transforming discovery chemistry.

Overall, we anticipate a gain of momentum for nanotechnology research and preview its future developments leveraged by ML concepts. The solutions we propose to the three outstanding challenges are realistic but surprisingly still not tackled by the research community. We expect the tight integration of computational technologies with robotics to result in a digital nanotechnology era that will see prototyping innovative and life-changing therapeutics done at a fraction of the time currently needed.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from FCT Portugal in the framework of PhD grant 2020.06638.BD (to D.P.S.), and the European Research Council grant agreement 848325 (J. Conde for the ERC Starting Grant). T.R. is an Investigador Auxiliar supported by FCT Portugal (CEECIND/00684/2018).

AUTHOR CONTRIBUTIONS

A.L., B.B.M., J. Conniot, and D.P.S. collected and analyzed data. J. Conde and T.R. conceived and supervised the research. All authors contributed to writing the manuscript and agreed on its final version.

DECLARATION OF INTEREST

J. Conde and T.R. are co-founders and shareholders of TargTex S.A.

1. Talebian, S., Rodrigues, T., das Neves, J., Sarmento, B., Langer, R., and Conde, J. (2021). Facts and Figures on Materials Science and Nanotechnology Progress and Investment. ACS Nano. https://doi.org/10.1021/acsnano.1021c03992.

2. Talebian, S., and Conde, J. (2020). Why Go NANO on COVID-19 Pandemic? Matter *3*, 598–601.

3. Schrurs, F., and Lison, D. (2012). Focusing the research efforts. Nat. Nanotechnol. *7*, 546–548.

4. Faria, M., Björnmalm, M., Thurecht, K.J., Kent, S.J., Parton, R.G., Kavallaris, M., Johnston, A.P.R., Gooding, J.J., Corrie, S.R., Boyd, B.J., et al. (2018). Minimum information reporting in bio-nano experimental literature. Nat. Nanotechnol. *13*, 777–785.

5. Conde, J. (2020). Above and Beyond Cancer Therapy: Translating Biomaterials into the Clinic. Trends Cancer *6*, 730–732.

6. Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., et al. (2017). The ChEMBL database in 2017. Nucleic Acids Res. *45* (D1), D945–D954.

7. Öztürk, H., Özgür, A., Schwaller, P., Laino, T., and Ozkirimli, E. (2020). Exploring chemical space using natural language processing methodologies for drug discovery. Drug Discov. Today *25*, 689–705.

8. Tao, H., Wu, T., Aldeghi, M., Wu, T.C., Aspuru-Guzik, A., and Kumacheva, E. (2021). Nanoparticle synthesis assisted by machine learning. Nat. Rev. Mater. *6*, 701–716.

9. Hart, G.L.W., Mueller, T., Toher, C., and Curtarolo, S. (2021). Machine learning for alloys. Nat. Rev. Mater. *6*, 730–755.

10. Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Mach. Learn. Sci. Technol. *1*, 045024.