Mini Review

BIOMEDICAL
Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2026.66.010288

Computational Challenges in Populating Data Deserts in Biomedical Research via Simulation

L Ridgway Scott*

The University of Chicago, Emeritus, USA

***Corresponding author:** L Ridgway Scott, The University of Chicago, Emeritus, USA

ARTICLE INFO

Received: June 26, 2026

Published: July 06, 2026

Citation: L Ridgway Scott. Computational Challenges in Populating Data Deserts in Biomedical Research via Simulation. Biomed J Sci & Tech Res 66(1)-2026. BJSTR. MS.ID.010288.

ABSTRACT

We discuss three different technical domains in biomedical research where simulation can be used to populate data deserts, all of which occur at three distinct physical scales. Moreover, we describe significant computational challenges in doing so for more demanding applications. In the case of fluid flow, we give precise Reynolds numbers that demarcate the current boundary between doable simulations and ones that require some advances in technology. The term "data desert" appears in many fields, including medicine [1,2], media research [3], meteorology research [4], as well as in many areas of technology [5]. These are parts of data space that are sparsely populated. This often limits the ability of AI to make reliable predictions [6]. On the other hand, there are many areas in which data is plentiful. Just picking fluid dynamics as an example, important technological phenomena are well studied experimentally [7] and computationally [8,9]. But there are many other areas of fluid mechanics where data is sparse. The purpose of this paper is to explore the use of simulation to populate data deserts.

In many areas, this approach does not work. But in areas in which there is a mathematical model for the phenomena of interest, this approach can potentially be effective. But it is critical that the simulation technology is reliable, which often requires advances in numerical analysis [10], algorithm design [11], and software development [12]. Often, extensive comparison of simulation techniques provides valuable guidance [9]. Moreover, the equations used to model the phenomena may have to be chosen carefully. Reduced models are frequently used which may not be sufficient. Our goal is not to settle all possible questions. Rather it is to raise issues that may lead to more reliable simulations, and thus to pose open questions that can lead to future developments. Undoubtedly, all of this will require continued advances in high performance (parallel) computing [13]. We have chosen three areas of biomedical research at different spatial and temporal scales to indicate how pervasive the issue of data deserts is. All three of these areas have significant impact outside of biomedical research, but for lack of space we do not describe them.

What is a Data Desert?

We give a simple example. There is a lot of discussion on the web about what life can be like if you have $1,000,000 in retirement savings. One might ask: how many people have saved $1,000,000 by retirement? Google AI says that there are 497,000 such Americans. But if you ask: how many people have saved $10,000,000 by retirement, the response is murkier. Data is given for $2,000,000 and $3,000,000, but "the number of individuals with $10 million in dedicated retirement accounts is a fraction of that figure" is all that Google AI tells us. We are left to wonder what that fraction might be. One could extrapolate the figures for k million dollars for k = 1, 2, 3 to k = 10, but it is not at all clear what extrapolation formula to use. The number for k = 10 is probably still not small, but so far this number is not in the collective data base. Of course, if it were needed for some reason, one can imagine it could be estimated accurately. But at the moment this is a data desert. It is likely the case that, although we are currently overwhelmed by massive data sets, the known parts of the data universe is probably only a tiny fraction of what could be known. Think of a fractal set of dimension ϵ for ϵ close to zero.

Can AI Extrapolate where there is no Data?

Yes, but not reliably. The example of retirement earnings is typical. If we knew the wealth distribution, we might be able to discern an extrapolation formula. But then there could be the issue of predicting extreme events [14].

DOI: 10.26717/BJSTR.2026.66.010288

Thus we limit discussion here to data interpolation, not data extrapolation. That is, we assume that data is abundant enough to make reliable predictions. When data is sparse, we propose simulation as a way to populate data space. Of course, this is limited to situations where reliable models are available. But we now consider three such cases.

Protein Modeling

Protein interactions provide a basis for drug discovery [15]. One of the greatest achievements of AI was recognized by a Nobel Prize in 2024 for protein folding. This accomplishment is based on vast data bases of protein structure solutions. These data bases have been used extensively to make predictions of protein behavior in many areas [16]. On the other hand, these data bases are sparse for portions of proteins which are not well structured, dubbed plastic regions in [5]. It was proposed in [5] that simulations could be used to predict possible structures that could be precipitated by a binding event. For small plastic regions, there is reason to expect that this approach could be successful. In [5], several simulation methodologies were suggested, but there are models at different levels of fidelity. At the highest level, quantum mechanics would be used, but this may be computationally prohibitive even for a few residues. Moreover, even classical quantum mechanics is being questioned in some circles [17]. An intermediate level would be Quantum Mechanics/Molecular Mechanics (QM/MM) methods, which were recognized by a Nobel Prize in 2013. Standard molecular modeling (MM) assumes properties of molecular interactions that may be inaccurate (incorrect physics) in some situations. The competition, namely experimental structure determination [6], uses real physics, not approximations. That is, the experiments determine behavior that may be difficult to simulate.

It is possible to compute the solution of quantum mechanics equations for small systems accurately [18-20]. But at the moment, such technology has not yet been extended to larger molecular systems.

Turbulence Modeling

Fluid mechanics governs many medical issues in the human body. Cardiovascular disease is perhaps the most well known [21,22], but fluid flow in the brain [23] is also a very active area of research, as well as in other organs [24]. Everyone who has flown on a commercial aircraft is aware of the word turbulence. This word is used to explain the bumpy ride that often occurs. But it turns out that turbulence is not a precisely defined concept. We know it when we see it, but there are many different fluid flows that could be described as turbulent. One such event occurs with flow in arteries [21,22]. Since commercial and private aircraft fly in large numbers everyday, it may seem counter-intuitive that there is a data desert here. But the problem is that instrumenting a commercial aircraft to measure the relevant data would be complicated. Unlike the case of molecular modeling in the previous section, the model equations for fluid flow are well established: the Navier–Stokes equations [25]. On the other hand, solving these equations accurately is challenging. Reduced equations, such as Reynolds-Averaged Navier–Stokes equations (RANS) [26] are popular, but inaccurate. Simulation of fluid flow has advanced substantially in the past fifty years. We now know how to identify reliable simulation technologies for fluid flow, but many important concepts are still in development.

For example, a basic benchmark in fluid dynamics is flow past a cylinder. For low speed flows (low Reynolds numbers), this can be done with sufficiently sophisticated techniques [9,27] for two-dimensional flows, but this benchmark shows the flaws in many simulation algorithms. For three-dimensional flows, the state of the art is more rudimentary. Thus it may take some time to fully populate the turbulence data desert. To provide a more complete picture, we focus on flow past a cylinder to give a sense of where simulation technology stands today. More precisely, we begin with a cylinder of infinite length and uniform flow (at infinity, or at a large distance upstream from the cylinder). We assume the flow direction is perpendicular to the cylinder axis. We will describe how the character of the flow changes as the speed is increased, but for precision we will use the Reynolds number as a descriptor for flow speed. The Reynolds number R is defined by $R = UL/\nu$ where L is the cylinder diameter, U is the flow speed at infinity, and ν is the fluid kinematic viscosity.

For R less than about 50, the flow is steady, and up-down symmetric, where we think of the plane perpendicular to the cylinder axis as the x, y plane, with z denoting the third dimension along the cylinder axis. More precisely, the velocity $\mathbf{u}$ satisfies $\mathbf{u}(x, y, z, t) = (u(x, y, z), v(x, y, z), 0)$ with symmetry $u(x, y, z) = u(x, -y, z)$ and $v(x, y, z) = -v(x, -y, z)$. Around $R = 50$, the flow becomes unsteady, but still planar (the component of $\mathbf{u}$ in the z-direction is zero). What is observed is the Karman vortex street [28], where vortices in the x, y plane are generated alternately from the top and bottom of the cylinder. The drag on the cylinder has been measured extensively experimentally [7] as well as computed via simulations. One measure of the flow is the drag-lift phase diagram [29]. This phase diagram remains periodic up to $R = 1000$, but it becomes chaotic by $R = 2000$ [9,30]. But notably, many simulations get this wrong, including [29].

Simulations in two dimensions are of significant industrial value for R up to a few thousand [31], but by $R = 10,000$ significant three-dimensional effects appear [32]. Beyond that, planar flow breaks down. It is suggested [33] that at some point the two dimensional vortices turn ninety degrees and instead a new type of vortex appears whose axis is parallel with the flow direction, instead of parallel with the cylinder axis as occurs for low Reynolds numbers. This has not been independently validated but it makes sense; if the lift oscillation continued to grow [34] then many cylinders (light and telephone poles, for example) would fall down. If the vortices turn 90 degrees, lift oscillation is essentially eliminated. To summarize the state-of-the-art, we are now able to reliably compute flow past a cylinder for Reynolds numbers up to a few thousand. On the other hand, many older

DOI: 10.26717/BJSTR.2026.66.010288

simulation schemes break down even before R = 1000. This means that doing reliable simulations for R > 10, 000 are a current challenge. But twenty years ago, reliable simulations for R = 1000 were not available, so we can project that advances in algorithms, software, and hardware will allow simulations for significantly higher Reynolds numbers.

Nonbonded Forces

Molecular modeling using quantum mechanics (Schrodinger's equation) reached a pinnacle with the 1998 Nobel Prize awarded in part for the development of the Gaussian code. That code has been widely used to predict structures of bonded molecular interactions. But nonbonded interactions, between molecules that are separated by several Angstroms, are more difficult since the approximation of functions in Gaussian are Gaussians, which decay too fast. Nonbonded interactions are critical components of molecular modeling codes [35-37]. Yet their approximation is often ad hoc. It is possible to approximate some nonbonded interactions (van der Waals forces) by solving the high-dimensional Schrodinger equation [18,20]. But so far, this has been carried out only for very simple systems (two hydrogen atoms).

Conclusion

We have indicated three different technical domains where simulation can be used to populate data deserts. Moreover, we have described significant challenges in doing so for more demanding applications. In the case of fluid flow, we have given precise Reynolds numbers that demarcate the current boundary between doable simulations and ones that require some advances in technology.

References

1. Behnood Bikdeli, Joseph S Ross, Harlan M Krumholz (2017) Data desert for inferior vena caval filters: limited evidence, supervision, and research. JAMA Cardiology 2(1): 3-4.
2. Robert A Winn, Katherine Y Tossas, Chyke Doubeni (2023) Commentary: Some water in the data desert: The cancer intervention and surveillance modeling network's capacity to guide mitigation of cancer health disparities. JNCI Monographs 2023(62): 167-172.
3. Lenka Waschkova Cısarova, Iveta Jansova, Jan Motal (2024) Delayed reflections: Media and journalism data deserts in the post-socialist czech republic. Media and Communication, p. 12.
4. Ross J Salawitch, Jessica B Smith, Henry Selkirk, Krzysztof Wargan, Martyn P Chipperfield, et al. (2025) The imminent data desert: The future of stratospheric monitoring in a rapidly changing world. Bulletin of the American Meteorological Society 106(3): E540-E563.
5. L Ridgway Scott, Ariel Fernandez (2024) Physics-informed AI infers drug-induced folding in target proteins. American Journal of Biomedical Science & Research 21(6): 647-650.
6. Neera Borkakoti, Janet M Thornton (2023) AlphaFold2 protein structure prediction: Implications for drug discovery. Current opinion in structural biology 78: 102526.
7. E R Relf (1914) Discussion of the results of measurements of the resistance of wires, with some additional tests on the resistance of wires of small diameter. Technical Report 102, Reports and Memoranda of the Aeronautical Research Council.
8. Ingeborg G Gjerde, L Ridgway Scott (2024) Verification and validation of cylinder drag: Pressure and stress approximations on curved boundaries. Advances in Computational Science and Engineering 2(3): 295-319.
9. Henry von Wahl, L Ridgway Scott (2024) Reliable chaotic transition in incompressible fluid simulations. Advances in Computational Science and Engineering 2(3): 202-221.
10. Ingeborg G Gjerde, L Ridgway Scott (2022) Nitsche's method for Navier-Stokes equations with slip boundary conditions. Mathematics of Computation 91(334): 597-622.
11. Brendan Keith, Thomas M Surowiec (2024) Proximal Galerkin: A structure-preserving finite element method for pointwise bound constraints. Foundations of Computational Mathematics 26: 385-481.
12. A Logg, KA Mardal, G Wells (2012) Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book. Springer-Verlag New York Inc.
13. L R Scott, T W Clark, B Bagheri (2005) Scientific Parallel Computing. Princeton University Press.
14. M Ghil, Pascal Yiou, Stephane Hallegatte, BD Malamud, P Naveau, et al. (2011) Extreme events: dynamics, statistics and prediction. Nonlinear Processes in Geophysics 18(3): 295-350.
15. Susan Jones, Janet M Thornton (1996) Principles of protein-protein interactions. Proceedings of the National Academy of Sciences 93(1): 13-20.
16. L Ridgway Scott, Ariel Fernandez (2017) A Mathematical Approach to Protein Biophysics. Springer Verlag.
17. Claes Johnson (2026) Quantum mechanics as multiphase 3D continuum mechanics.
18. Eric Cances, L Ridgway Scott (2018) van der Waals interactions between two hydrogen atoms: The Slater-Kirkwood method revisited. SIAM J Applied Mathematics 50(1): 381-410.
19. Eric Cances, Rafael Coyaud, L Ridgway Scott (2023) van der Waals interactions between two hydrogen atoms: The next orders. Communications in Mathematical Sciences 21(4): 915-948.
20. Alexander Veit, L Ridgway Scott (2017) Using the tensor-train approach to solve the ground-state eigenproblem for hydrogen molecules. SIAM J Sci Computing 39(1): B190-B220.
21. Nanna Berre, Gabriela Castro, Henrik Kjeldsberg, Rami Masri, Ingeborg Gjerde (2022) A computational study of flow instabilities in aneurysms. In Computational Physiology: Simula Summer School 2021-Student Reports, p. 63-75.
22. Alessandro Contri, Christina Taylor, Justin Tso, Ingeborg Gjerde (2023) Inducing flow instabilities in aneurysm geometries via the reynolds–orr method. In Computational Physiology: Simula Summer School 2022-Student Reports, p. 79-89.
23. Vegard Vinje, Geir Ringstad, Erika Kristina Lindstrøm, Lars Magnus Valnes, Marie E Rognes, et al. (2019) Respiratory influence on cerebrospinal fluid flow–a computational study based on long-term intracranial pressure measurements. Scientific Reports 9(1): 9732.
24. M J Lighthill (1972) Physiological fluid dynamics: a survey?. Journal of Fluid Mechanics 52(3): 475-497.
25. L D Landau, E M Lifshitz (1959) Fluid Mechanics. Pergamon Press.
26. Phillippe R Spalart (2009) RANS modelling into a second century 23(4):

291-293.

27. Henry von Wahl, Leo Rebholz, L Ridgway Scott (2026) Benchmark stress tests for flow past a cylinder at higher Reynolds numbers using EMAC. International Journal of Numerical Methods for Heat and Fluid Flow 36(7): 2444:2459.
28. Ming Hsun Wu, Chih Yung Wen, Ruey Hor Yen, Ming Cheng Weng, An Bang Wang (2004) Experimental and numerical study of the separation angle for flow around a circular cylinder at low Reynolds number. Journal of Fluid Mechanics 515: 233-260.
29. L Ridgway Scott, Rebecca Durst (2024) Chaotic dynamics of two-dimensional flows around a cylinder. Physics of Fluids 36(2): 024118.
30. D Durante, C Pilloton, A Colagrossi (2022) Intermittency patterns in the chaotic transition of the planar flow past a circular cylinder. Physical Review Fluids 7(5): 054701.
31. L Ridgway Scott (2025) Industrial applications of lift and drag forces in chaotic flow. In Matteo Giacomini, Simona Perotto, and Gianluigi Rozza, editors, Emerging Technologies in Computational Sciences for Industry, Sustainability and Innovation: M2P 2023, Lecture Notes in Computational Science and Engineering 146: 349-358.
32. Michio Hayakawa, Fazle Hussain (1989) Three-dimensionality of organized structures in a plane turbulent wake. Journal of Fluid Mechanics 206: 375-404.
33. Johan Hoffman, Johan Jansson, Claes Johnson (2016) New theory of flight. Journal of Mathematical Fluid Mechanics 18(2): 219-241.
34. L Ridgway Scott, Henry von Wahl (2026) Lift data for flow around a cylinder. Tbd.
35. C L Brooks, M Karplus, B M Pettitt (1988) Proteins: A theoretical perspective of dynamics, structure, and thermodynamics 71: 280.
36. S T Wlodek, T W Clark, L R Scott, J A McCammon (1997) Molecular dynamics of acetylcholinesterase dimer complexed with tacrine. J Am Chem Soc 119: 9513-9522.
37. T W Clark, J A McCammon, L R Scott (1992) Parallel molecular dynamics. In: J. Dongarra (Edt.)., Proc Fifth SIAM Conf on Parallel Proc for Sci Comp, pp. 338-344.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2026.66.010288

L Ridgway Scott . Biomed J Sci & Tech Res