more than the sum of the parts. In previous Cas12k–RNA–DNA structures[13,14], the DNA was only partially paired with the RNA — but the new work shows that pairing is complete when TniQ, TnsC and S15 are present. The latest structures also show that the strand of the DNA duplex in closest contact with TnsC in the full assembly is different from what was seen in previous structures of isolated TnsC–DNA complexes[14,15].

Mutual interactions between TnsC and TnsB are a key aspect of transposon targeting[18]. The TnsC filament is trimmed down to size by TnsB, to bring the transposase into close proximity with the target DNA sequence specified by the guide RNA. In turn, the catalytic activity of TnsB (unlike many transposases that do their job alone) depends on interactions with TnsC.

A surprising observation from Park and colleagues' structures was that the number of TnsC molecules in the minifilament was variable (12 or 13). This flexibility suggests how a system that can so accurately target a particular DNA sequence can also show 5–10 base pairs of variability in the insertion site selected. Such a feature might reflect an adaptation to the natural battlefield, enabling small variations in CAST insertion that minimize the harm of its addition to the host. Efforts are now needed to describe the dynamics of TnsC assembly and disassembly, and to elucidate TnsB's exact role in it.

How is transposase activity controlled? One might have expected that TnsC would alter the conformation of TnsB to activate its catalytic core. Therefore, it is interesting that the TnsB structure Park *et al.* observed in the full complex is similar to that of TnsB without TnsC[16,17]. Moreover, the authors found no contacts between TnsC and the catalytic core of TnsB that could activate TnsB.

Clues to how TnsC activates TnsB might be found by considering a related transposase called MuA. For both TnsB and MuA, the target DNA must strongly bend to fit into the active pocket of the transposase. Making such a strong bend in DNA is energetically difficult, and factors that stabilize bent target DNA enhance MuA transposition[19]. Perhaps TnsC activates transposition, at least in part, by bringing bent target DNA to TnsB. In fact, TnsB might not be able to bind to natural target DNA in the absence of TnsC, which tethers it to DNA. Moreover, in the structure reported by Park *et al.*, the TnsC minifilament is positioned to contact target DNA on both sides of TnsB, perhaps to stabilize its bent form. Finally, the authors observed that interactions between TnsB and TnsC caused a disordered segment of TnsB to fold and dock in a way that might help to stabilize bent target DNA.

Together, these two papers highlight how CRISPR and transposons — which conventionally oppose one another — can join forces to strategically choose target locations for inserting large DNA payloads. The molecular views provided will help genome engineers to develop CAST-based gene-insertion systems with more-tailored target locations, and could help researchers to design variants of the system that exhibit less off-target activity.

The unexpected discovery that the bacterial S15 protein is an intrinsic *Sh*CAST component could prove to be a breakthrough in achieving large targeted insertions in eukaryotic genomes. Schmitz and colleagues show that S15's human relative does not promote *Sh*CAST activity, which could explain why previous attempts to adapt the system for use in eukaryotic cells have failed. The ability of bacterial S15 to boost *Sh*CAST activity in eukaryotes will surely be tested soon.

Finally, the CAST system could potentially be used as a blueprint to establish programmable targeting of simpler transposons. This could aid CRISPR-mediated gene insertion in diverse cells and organisms — an exciting prospect for future research.

Orsolya Barabas is in the Department of Molecular and Cellular Biology, University of Geneva, 1211 Geneva, Switzerland. **Phoebe A. Rice** is in the Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, Illinois 60637, USA.
e-mails: orsolya.barabas@unige.ch; price@uchicago.edu

1. Park, J.-U. *et al. Nature* **613**, 775–782 (2023).
2. Schmitz, M., Querques, I., Oberli, S., Chanez, C. & Jinek, M. *Cell* **185**, 4999–5010 (2022).
3. Hille, F. *et al. Cell* **172**, 1239–1259 (2018).
4. Knott, G. J. & Doudna, J. A. *Science* **361**, 866–869 (2018).
5. Ivics, Z. *et al. Nature Methods* **6**, 415–422 (2009).
6. Hickman, A. B. & Dyda, F. *Microbiol. Spectr.* **3**, MDNA3-0034-2014 (2015).
7. Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. *Proc. Natl Acad. Sci. USA* **114**, E7358–E7366 (2017).
8. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. *Nature* **571**, 219–225 (2019).
9. Strecker, J. *et al. Science* **365**, 48–53 (2019).
10. Vo, P. L. H. *et al. Nature Biotechnol.* **39**, 480–489 (2021).
11. Rubin, B. E. *et al. Nature Microbiol.* **7**, 34–47 (2022).
12. Halpin-Healy, T. S., Klompe, S. E., Sternberg, S. H. & Fernández, I. S. *Nature* **577**, 271–274 (2020).
13. Xiao, R. *et al. Mol. Cell* **81**, 4457–4466 (2021).
14. Querques, I., Schmitz, M., Oberli, S., Chanez, C. & Jinek, M. *Nature* **599**, 497–502 (2021).
15. Park, J.-U. *et al. Science* **373**, 768–774 (2021).
16. Park, J. U., Tsai, A. W.-L., Chen, T. H., Peters, J. E. & Kellogg, E. H. *Proc. Natl Acad. Sci. USA* **119**, e2202590119 (2022).
17. Tenjo-Castaño, F. *et al. Nature Commun.* **13**, 5792 (2022).
18. Skelding, Z., Sarnovsky, R. & Craig, N. L. *EMBO J.* **21**, 3494–3504 (2002).
19. Fuller, J. R. & Rice P. A. *eLife* **6**, e21777 (2017).

O.B. declares competing interests. See go.nature.com/3jqejfi for details.
This article was published online on 11 January 2023.

Catalysis

Machine learning classifies reaction mechanisms

Danilo M. Lustosa & Anat Milo

The study of how chemical reactions work is key to the design of new reactions, but relies on hard work and expert knowledge. A machine-learning tool has been developed that could change the way this challenge is approached. **See p.689**

The discovery of chemical reactions is influenced not only by how fast experimental data can be acquired, but also by how easily chemists can make sense of these data. Unravelling the mechanistic underpinnings of new catalytic reactions is a particularly intricate problem, often requiring expert knowledge of computational and physical organic chemistry. Nevertheless, it is important to study catalytic reactions because they represent the most efficient chemical processes. On page 689, Burés and Larrosa[1] report a machine-learning model that classifies the mechanisms of catalytic reactions on the basis of the time-course signatures of the reactions. This method could streamline the investigation of reaction mechanisms and requires minimal experimental effort.

The determination of catalytic-reaction mechanisms involves collecting a plethora of clues about how starting materials come together and interact with a catalyst and each other to form products. One of the most powerful techniques for drawing hypotheses from experimental data is to analyse the consumption of starting materials and the formation of products over time, a process called kinetic analysis[2]. These rates of reactant decay and product formation are best described by equations known as rate laws.

The basic premise behind these laws is that the rate of a chemical transformation depends on the concentration of reagents, the number of reagent molecules involved in each step of the reaction, and how much energy is needed to transform the reactants to transient intermediates that then form the products. Disentangling complex rate laws can lead to

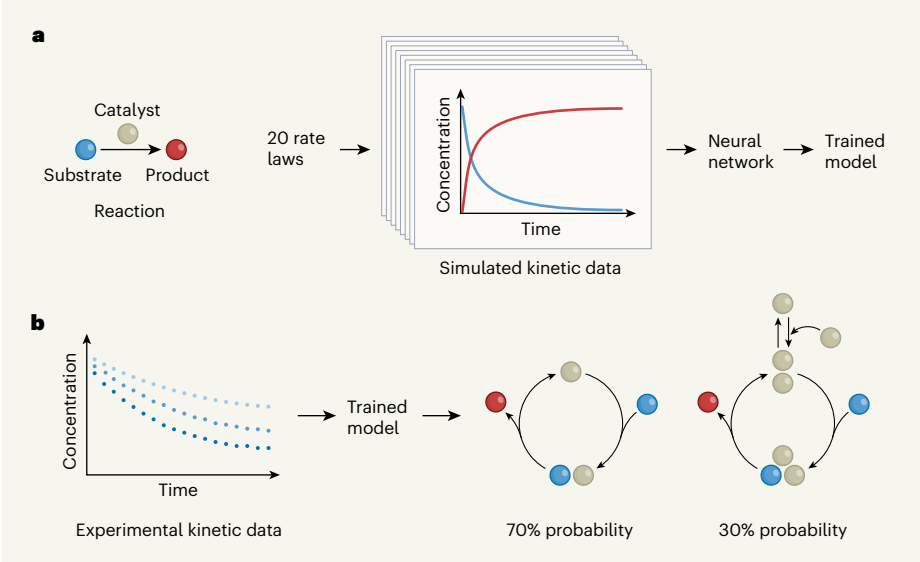

Figure 1 | Development of a model for the kinetic analysis of chemical reactions. a, Burés and Larrosa[1] considered a general set of reactions in which a substrate molecule is converted into a product by a catalyst. The authors identified 20 possible mechanisms, and worked out the rate law for each — an equation that describes the reaction kinetics (the rates at which the substrate is consumed and the product is formed). The rate laws were solved to generate millions of simulated kinetic profiles, and the solutions were fed into a machine-learning system (a neural network) to produce a trained model of reaction kinetics. **b**, The model can process kinetic data from real reactions and accurately identify possible reaction mechanisms, assigning a probability for each. Here, the most likely mechanism is a simple process in which the substrate forms a complex with one molecule of the catalyst, and then produces a product; the less-likely mechanism involves the substrate forming a complex with a dimer of the catalyst.

a better understanding of which molecules are involved in each step of a catalytic cycle and the sequence in which these steps occur. Therefore, it is not surprising that mechanistic information has broad implications for the design of new reactions.

Chemists have been extracting mechanistic information from reaction rates for more than a century. A method still used today is to evaluate the initial rate of a reaction[3], focusing on the consumption of the first few per cent of starting material. This method is popular because, in most cases, the plots of reactant concentration against time are linear at the start of a reaction, and are therefore straightforward to analyse. Although insightful, this technique disregards changes in reaction rates and concentrations that occur throughout most of the time course.

In the past couple of decades, more-advanced methods have been developed that evaluate the concentrations of reaction components throughout the full course of a reaction[4]. These methods have been further facilitated by mathematical techniques that reveal the number of components involved in a reaction step (also known as the order of reaction components) from plots of reaction kinetics[5]. Such techniques will surely continue to provide great insights into chemical reactivity, but they are confined to analysing the order of reaction components, rather than providing a more holistic mechanistic hypothesis that describes the kinetic behaviour of a catalytic system.

Machine learning is revolutionizing the way that chemists solve problems, from designing molecules and routes to synthesizing them[6,7], to understanding reaction mechanisms[8]. Burés and Larrosa now bring this revolution to kinetic analysis through a machine-learning model that classifies reactions on the basis of simulated kinetic signatures of reactions.

The authors defined 20 classes of reaction mechanism, and worked out the rate law for each of them (Fig. 1). They then solved these equations to generate millions of simulations describing reactant decay and product formation. These simulated kinetic data were used to train a learning algorithm to identify the characteristic signatures of each mechanistic class. The resulting classification model uses kinetic profiles as input, including initial and temporal concentration data, and outputs the mechanistic class of a reaction.

"This approach for training the algorithm avoids the bottleneck of producing an immense amount of experimental kinetic data."

Burés and Larrosa's approach for training their algorithm avoids the bottleneck of producing an immense amount of experimental kinetic data, which are needed to feed data-hungry deep-learning systems — generating such a quantity of data in a laboratory would take years. Moreover, the simulated data are 'cleaner' for training purposes, because each kinetic profile is inherently associated with a specific mechanistic scenario.

The authors evaluated the trained model with a test set of simulated kinetic profiles, and demonstrated that it correctly assigns these profiles to a mechanistic class with 92.6% accuracy. The model performed well even when 'noisy' data were intentionally introduced, which implied that it would be useful for categorizing experimental data.

Finally, the authors benchmarked their model using several previously reported experimental kinetic profiles. The predicted mechanisms agreed closely with the conclusions from the earlier kinetic studies. In some cases, the model also identified mechanistic details that were not detected in the original work. For one challenging reaction, three very similar mechanistic classes were suggested by the model. However, the authors rightly say that this outcome is not a bug, but rather a feature of their model, because it suggests the need for specific further experiments to explore the mechanism.

In sum, Burés and Larrosa have developed a method that not only automates the lengthy process of deriving mechanistic hypotheses from kinetic investigations, but also enables kinetic analysis of challenging reaction mechanisms. Like any technological advance in data analysis, the resulting mechanistic classifications should be treated as hypotheses that require further experimental support. There is always a risk of misinterpreting kinetic data, but the ability of the algorithm to identify the correct reaction path (or paths) on the basis of a small number of experiments, with a high degree of accuracy, could persuade more researchers to attempt kinetic analysis. We therefore share the authors' enthusiasm that this approach can popularize and advance the inclusion of kinetic analysis in the pipeline of reaction development, especially as chemists become more comfortable with machine-learning algorithms.

Danilo M. Lustosa and **Anat Milo** are in the Department of Chemistry, Ben-Gurion University of the Negev, Be'er Sheva 84105, Israel.
e-mail: anatmilo@bgu.ac.il

1. Burés, J. & Larrosa, I. *Nature* **613**, 689–695 (2023).
2. Blackmond, D. G. *J. Am. Chem. Soc.* **137**, 10852–10866 (2015).
3. Michaelis, L. & Menten, M. L. *Biochem. Z.* **49**, 333–369 (1913).
4. Blackmond, D. G. *Angew. Chem. Int. Edn* **44**, 4302–4320 (2005).
5. Burés, J. *Angew. Chem. Int. Edn* **55**, 16084–16087 (2016).
6. Williams, L. W. *et al. ACS Cent. Sci.* **7**, 1622–1637 (2021).
7. Strieth-Kalthoff, F., Sandfort, F., Segler, M. H. S. & Glorius, F. *Chem. Soc. Rev.* **49**, 6154–6168 (2020).
8. Lustosa, D. M. & Milo, A. *ACS Catal.* **12**, 7886–7906 (2022).

The authors declare no competing interests.