Exemplary fluorescence-intensity images for origamis containing 12, 24 or 36 ATTO647N molecules indicated the homogeneity of the samples. Analysis of the photons per spot for all samples revealed a linear dependence of intensity on dye number (**Fig. 1a**) and excitation intensity (**Supplementary Fig. 1a**). The linear dependence on dye number is notable because samples with such a high density of fluorophores commonly exhibit self-quenching. A comparison with commercially available fluorescent beads (dark red FluoSpheres (Invitrogen) showed that DNA origamis with as few as 12 ATTO647N molecules are brighter than FluoSpheres and exhibit a more homogeneous intensity and fluorescence lifetime (**Supplementary Fig. 1b–e**). Labeled DNA origamis can therefore be used as brightness standards to quantify microscope sensitivities and relate emission intensities to absolute numbers of fluorescent molecules.

With the DNA origami platform, dyes can be positioned at distances from the Förster resonance energy transfer range[2] up to distances corresponding to the longest axis of the specific DNA origami. For a calibration standard that can be resolved by conventional diffraction-limited microscopes, we constructed a six-helix bundle that has a persistence length of ~2 μm (ref. 3) and placed two ATTO647N molecules at a contour-length distance of 386 nm (**Fig. 1b**). A confocal image shows the resulting dumbbell-shaped spots that were resolved by the diffraction-limited microscope. The histogram of the spot separations (**Supplementary Fig. 2**) yields a distance of 357 nm, demonstrating that robust standards exceeding diffraction-limited dimensions could be constructed from single DNA origami scaffolds.

STED was the first far-field super-resolution microscopy method that abandoned the diffraction barrier, and it is the most prominent representative of a family of super-resolution microscopy techniques based on targeted switching of molecules in a predetermined region[4]. We created a nanoruler for pulsed STED with a distance of 71 nm between two parallel lines of 12 ATTO647N molecules each. The lines were not resolved in a confocal microscope with 5% STED beam intensity (**Fig. 1c**). An increase of the doughnut-shaped STED beam intensity to 50% of its maximum caused the spots to shrink because of stimulated emission quenching in the outer parts of the laser focus, and at 100% STED beam intensity, corresponding to 110 mW, the two lines were well resolved (**Fig. 1c**). Analysis of the 71-nm ruler, a 44-nm ruler and an 80-nm ruler for continuous-wave STED are presented in **Supplementary Figures 3** and **4**.

An alternative approach for super-resolution imaging exploits the successive localization of single blinking or photoactivatable molecules[5]. We broadened our initial experiments on single-molecule localization on DNA origami[6] to measure <100-nm distances in different spectral ranges (**Fig. 1d** and **Supplementary Figs. 5** and **6**).

To show that DNA origami standards can cover the full range from diffraction-limited to molecular dimensions, we constructed DNA origami rectangles with two ATTO647N molecules separated by a distance of 6, 12 or 18 nm (**Fig. 1e**). For these standards, we used successive photobleaching and analyzed fluorescence transients from identified spots with respect to photobleaching steps. The individual molecules were localized in reverse order of photobleaching, and the intensity distribution of the second molecule was subtracted from that of the first part of the transient (**Fig. 1e**). The corresponding localization map shows a clear separation of the two dye molecules. The histogram of localizations reveals a distance of 5.7 nm well in accordance with the expected value of 6 nm, assuming an interhelical separation of 3 nm (ref. 1). All three distances (d) could be well reproduced with values of $d_1 = 5.8 \pm 2.9$ nm, $d_2 = 10.7 \pm 1.8$ nm and $d_3 = 18.3 \pm 5.7$ nm (mean ± s.d., $n = 30$) (**Fig. 1e**).

DNA origami–based fluorescence standards have matured into ready-to-use validation samples. After being covered with a polymer layer, the origami standards can be transported and stored for 12 months (**Supplementary Figs. 7** and **8**). We propose DNA origamis as a standard platform to test and prove abilities of new super-resolution techniques as well as for everyday use to distinguish instrument-specific from sample-specific error sources in fluorescence imaging.

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nmeth.2254.

ACKNOWLEDGMENTS

We thank V. Schüller, R. Schreiber, E. Pibiri, I. Burkhardt, R. Jungmann, P. Nickels, T. Liedl, B. Lalkens and D. Grohmann for help with experiments and fruitful discussions. We are grateful to W. Fouquet, J. Sieber and Leica Microsystems for STED measurements. Financial support by the Biophotonics IV program of the Federal Ministry of Education and Research (BMBF)/Association of German Engineers (VDI)(13N11461) and the German Research Foundation (DFG Ti329/6-1) is gratefully acknowledged.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available at http://www.nature.com/doifinder/10.1038/nmeth.2254.

Jürgen J Schmied[1,3], Andreas Gietl[1,3], Phil Holzmeister[1,3], Carsten Forthmann[1], Christian Steinhauer[2], Thorben Dammeyer[1] & Philip Tinnefeld[1]

[1]Institute for Physical and Theoretical Chemistry, Braunschweig University of Technology, Braunschweig, Germany. [2]STS Nanotechnology UG, Freising, Germany. [3]These authors contributed equally to this work.
e-mail: p.tinnefeld@tu-braunschweig.de

1. Rothemund, P.W. *Nature* **440**, 297–302 (2006).
2. Stein, I.H., Schüller, V., Böhm, P., Tinnefeld, P. & Liedl, T. *Chemphyschem* **12**, 689–695 (2011).
3. Liedl, T., Högberg, B., Tytell, J., Ingber, D.E. & Shih, W.M. *Nat. Nanotechnol.* **5**, 520–524 (2010).
4. Hell, S.W. *Science* **316**, 1153–1158 (2007).
5. Betzig, E. *et al.* **313**, 1642–1645 (2006).
6. Steinhauer, C., Jungmann, R., Sobey, T.L., Simmel, F.C. & Tinnefeld, P. *Angew. Chem. Int. Ed. Engl.* **48**, 8870–8873 (2009).

Flaws in evaluation schemes for pair-input computational predictions

To the Editor: Computational prediction methods that operate on pairs of objects by considering features of each (hereafter referred to as pair-input methods) have been crucial in many areas of biology and chemistry over the past decade. Among the most prominent examples are protein-protein interaction (PPI)[1], protein-drug interaction[2,3], protein-RNA interaction[4] and drug indication[5] prediction methods. A sampling of more than 50 published studies involving pair-input methods is provided in **Supplementary Table 1**. Here we demonstrate that the paired nature of inputs has significant, though not yet widely perceived, implications for the validation of pair-input methods.

The effects that the paired nature of inputs has on the cross-validation of pair-input methods can be seen in the following example. Proteochemometrics modeling[2], a computational

methodology for predicting protein-drug interactions, uses a feature vector for a chemical and a feature vector for a protein receptor to predict the binding between them[2]. In this case, a test pair may share either the chemical or protein component with some pairs in a training set; it may also share neither. We found that pair-input methods perform much better for test pairs that share components with pairs in a training set than for those that do not. As a result, it is necessary to distinguish test pairs on the basis of whether they share components with pairs in a training set when evaluating performance.

A test set used to estimate predictive performance may be dominated by pairs that share components with training pairs in the training set, yet such pairs may be a minority on the population level. In this case, a predictive performance estimated on the test set may be impressive, yet it will likely fail to generalize to the population level. Indeed, this issue has been previously recognized by some researchers[6] (**Supplementary Table 1**). However, it has been overlooked by many, and cross-validations for pair-input methods usually do not distinguish test pairs on the basis of this component-level overlap criterion (**Supplementary Table 1**).

To illustrate the issue, we consider PPI prediction methods with a toy example (**Fig. 1**), in which the protein space is composed of nine proteins and a training set consists of four positive and four negative protein pairs. This training set is used to train a PPI prediction method, which is in turn applied to a set of 28 test pairs. How well would the trained method perform on the 28 test pairs? To determine this, one usually performs a cross-validation on the training set. A temporary training set is prepared by randomly picking some pairs (**Fig. 1**), and the rest serve as a temporary test set from which predictive accuracy can be measured. This cross-validated predictive performance is then implicitly assumed to hold for the full space of 28 test pairs. The paired nature of inputs leads to a natural partitioning of the 28 test pairs into three distinct classes (C1–C3, **Fig. 1**): C1 has test pairs sharing both proteins with the training set, C2 has test pairs sharing only one protein with the training set, and C3 has test pairs sharing neither protein with the training set.

To demonstrate that the predictive performance of pair-input methods differs significantly for distinct test classes, we performed computational experiments using large-scale yeast and human PPI data that mirror the toy example (**Supplementary Methods**). For all seven PPI prediction methods (M1–M7, chosen to be a representative set of algorithms; **Supplementary Methods**), the predictive performances for the three test classes differ significantly (**Supplementary Table 2**). The differences are not only statistically significant (**Supplementary Table 3**) but in many cases also numerically large. M1–M4 are support vector machine (SVM)-based methods, M5 is based on the Random Forest algorithm and M6 and M7 are heuristic methods. Thus, regardless of core predictive algorithms, significant differences for the three distinct test classes are consistently observed. These differences arise partly from the learning of differential representation of components among positive and negative training examples (**Supplementary Discussion** and **Supplementary Table 4**).

In a typical cross-validation for pair-input methods, available data are randomly divided into a training set and a test set, without regard to the partitioning of test pairs into distinct classes.

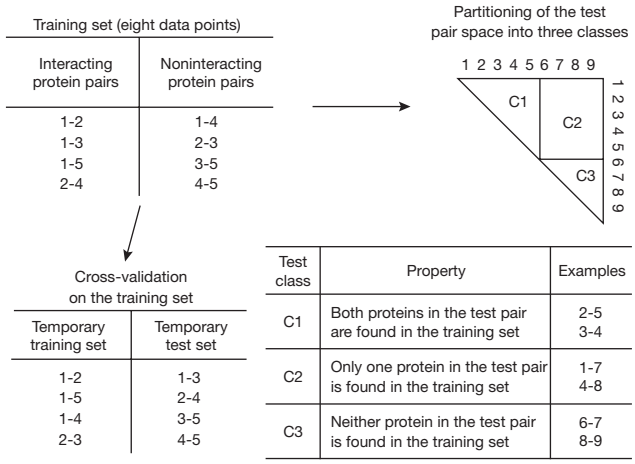

Figure 1 | Toy example (nine proteins) demonstrating prediction of protein-protein interactions and showing the shortcomings of a typical cross-validation scheme. The training and test sets for a PPI prediction method are shown. The paired nature of inputs leads to a natural partitioning of the 28 test pairs into three distinct classes (C1, C2 and C3). Cross-validation on the training set is typically done by randomly dividing the eight training pairs into a temporary training set and a temporary test set as shown, without regard to the partitioning of test pairs into distinct classes, and can therefore misrepresent overall predictive performance.

How representative would such randomly generated test sets be of full populations? To answer this question, we performed the typical cross-validation using the same yeast and human PPI data. Not surprisingly, the C1 class accounted for more than 99% of each of the test sets generated for typical cross-validations, and accordingly the cross-validated predictive performances closely match those for the C1 class (**Supplementary Table 2**). In contrast, within the full population (that is, the set of possible human protein pairs), the C1 class represents only a minority of cases: 21,946 protein-coding human genes[7] imply 240,802,485 possible human protein pairs. According to HIPPIE[8], a meta-database integrating ten public PPI databases, the space of C1-type human protein pairs accounts for only 19.2% of these cases, compared with 49.2% and 31.6%, respectively, for the C2 and C3 classes. Hence, the C1 class is far less frequent at the population level than for typical cross-validation test sets, and performance estimates obtained by a typical cross-validation should not be expected to generalize to the full population level.

In summary, computational predictions—whether pair-input or not[9,10]—that are tested by cross-validation on nonrepresentative subsets should not be expected to generalize to the full test populations. A unique aspect of pair-input methods, as compared with methods operating on single objects, is that one additionally needs to take into account the paired nature of inputs. We have demonstrated that (i) the paired nature of inputs leads to a natural partitioning of test pairs into distinct classes, and (ii) pair-input methods achieve significantly different predictive performances for distinct test classes. We note that if one is only interested in the population of C1 test pairs, then typical cross-validations using randomly generated test sets are acceptable, although this limitation should then be noted. For general-purpose pair-input methods, however, it is imperative to distinguish distinct classes of test pairs, and we propose that predictive performances should be reported separately for each distinct test class. In the case of PPI prediction methods,

three independent predictive performances should be reported (**Supplementary Table 2**). In the case of protein-drug interaction prediction methods, one should report four independent predictive performances, as either the protein or drug component of a test pair might each be shared with pairs in training data.

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nmeth.2259.

ACKNOWLEDGMENTS

We thank W.S. Noble, A. Ben-Hur, J.-P. Vert and V. Helms for stimulating discussions and critical comments on the manuscript. This work was supported by grants to E.M.M. from the US National Institutes of Health, the US Army (58343-MA), Cancer Prevention and Research in Texas and the Welch (F1515) Foundation. Y.P. acknowledges financial support from the Deutsche Forschungsgemeinschaft (DFG-Forschungsstipendium).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Yungki Park & Edward M Marcotte

Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA.
e-mail: yungki@mail.utexas.edu or marcotte@icmb.utexas.edu

1. Shen, J. *et al. Proc. Natl. Acad. Sci. USA* **104**, 4337–4341 (2007).
2. Wikberg, J.E. & Mutulis, F. *Nat. Rev. Drug Discov.* **7**, 307–323 (2008).
3. Yabuuchi, H. *et al. Mol. Syst. Biol.* **7**, 472 (2011).
4. Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G.G. *Nat. Methods* **8**, 444–445 (2011).
5. Gottlieb, A., Stein, G.Y., Ruppin, E. & Sharan, R. *Mol. Syst. Biol.* **7**, 496 (2011).
6. Vert, J.-P. & Yamanishi, Y. *Advances in Neural Information Processing Systems* vol. 17 (eds. Saul, L., Weiss. Y. & Bottou, L.) 1433–1440 (MIT Press, Cambridge, Massachusetts, USA, 2005).
7. Flicek, P. *et al. Nucleic Acids Res.* **39**, D800–D806 (2011).
8. Schaefer, M.H. *et al. PLoS ONE* **7**, e31826 (2012).
9. Tropsha, A. & Golbraikh, A. *Curr. Pharm. Des.* **13**, 3494–3504 (2007).
10. Olah, M., Bologa, C. & Oprea, T.I. *J. Comput. Aided Mol. Des.* **18**, 437–449 (2004).