Supplementary Methods

An extension for handling the distribution of \( p(S|D) \) with the multiple copy model

\( p(S|D) \) might vary from one domain to another. It is indeed possible to use a different \( p(S|D) \) value for each domain, instead of an average value as in Eq. 7, \( \overline{p(S|D)} = \frac{\sum_{i=1}^{M} p_i(S|D)}{M} \cdot p(S|D) \) for domain \( i \) is denoted by \( p_i(S|D) \). Assuming \( p(D) \) is constant, the difference between

\[
E_{\text{domain}}(p_i(S|D)) = p(D) \sum_{i=1}^{M} \left[ 1 - \{1 - p_i(S|D)\}^N \right]
\]  
(A1)

and

\[
E_{\text{domain}}(p(S|D)) = p(D) M \left[ 1 - \{1 - p(S|D)\}^N \right]
\]  
(A2)

is derived as follows. Using

\[
g(p,N) = 1 - (1 - p)^N \\
= 1 - \left[ \{1 - Np + N C_2 p^2 - N C_3 p^3 + \cdots + (-1)^{N-1} N C_{N-1} p^{N-1} + (-1)^N p^N \} \right] \\
= Np - N C_2 p^2 + N C_3 p^3 + \cdots - N (-1)^{N-1} N C_{N-1} p^{N-1} - (-1)^N p^N
\]

(A3)

with \( p \) as \( p(S|D) \) and dividing \((A1 - A2)\) by constant \( p(D) \), we obtain

\[
\frac{E_{\text{domain}}(p_i(S|D)) - E_{\text{domain}}(p(S|D))}{\text{constant } p(D)} = e_{\text{domain}}(p_i(S|D)) - e_{\text{domain}}(p(S|D)) \\
= \frac{1}{N+r} \left[ \frac{1}{M} \sum_{i=1}^{M} \left[ 1 - \{1 - p_i(S|D)\}^N \right] - \left[1 - \{1 - p(S|D)\}^N \right] \right] \\
= \frac{1}{N+r} \left[ g(p,N) - g(p,N) \right].
\]

(A4)
The equation (A4), which yields the difference of the expected number of obtaining different domains between extended handling of the distribution and our approximated model, becomes zero if \( N \) goes to infinity. The difference is also 0 if \( N = 1 \). Furthermore, simulations using our experimentally observed distribution \( \{p_i\} \) in Fig. 4a, excluding the fragments in the subset \( D^C \) (Fig. 1c), resulted in \( \frac{p(S|D)}{p(S|D)} = 0.71 \). Substituting this into A4, revealed that \( e_{\text{domain}}(p_r(S|D)) \) was smaller than \( e_{\text{domain}}(p(S|D)) \) by 0.044, 0.030, 0.0033, and 0.0030 for \((N, r) = (2, 0), (2, 1), (10, 0), \) and \((10, 1), \) respectively, which correspond to errors of 3 to 10% for the value of \( e_{\text{domain}}(p(S|D)) \).

**Basic properties of \( P_N \)**

In practical terms, the values computed using Eqs. 6 and 7 are overlapping (for most values, within a few percent error; see also Fig. 5). The virtue of Eq. 6 is its simplicity, and in its continuous form it can be readily differentiated. \( N_{\text{optimum}} \), which is the value of \( N \) that maximizes \( E_{\text{domain}} \) is thus obtained by differentiating Eq. 6 with respect to \( N \) and equating it to 0, which yields

\[
\exp(aN) = aN + ar + 1,
\]

where \( a = -\log[1 - p(S|D)] = -\log[1 - f / F] \geq 0 \). The left-hand side is an exponentially increasing function of \( N \), and the right-hand side is a linearly increasing function of \( N \) with an ordinate of \( 1+ar \). This implies that Eq. A5 has only a single solution: if the experimental set-up cost \( (r) \) is nil, then the solution to Eq. A5 is \( N=0 \) because \( ar = 0 \). Thus, if no set-up costs are incurred upon the examination of a new domain, then the most effective way is to test one fragment per domain (since zero is not a realistic solution). This is clearly illustrated in Fig. 5a, which shows that \( P_N \) is a monotone decreasing function of \( N \).

One peculiar situation arises in the single copy model for \( f=1 \) and \( r=0 \) (Eq. 7, with no set-up cost). Eq. 7 becomes

\[
E_{\text{domain}}(f=1, r=0) = \frac{\text{constant}}{N} p(D)\left[1 - \frac{(F-1)(F-N)}{(F-1-N)F!}\right] = \frac{\text{constant}}{N} p(D)\frac{N}{F}
\]
\[ = \frac{\text{constant}}{F} \cdot p(D) \]

in which, surprisingly, the dependence on \(N\) vanishes. Although this is a very specific case, it is noteworthy that this situation is not reproduced with Eq. 6, where multiple copies of the same fragments can be generated. This difference arises because in the single copy model, the chances of producing a soluble fragment improve with increasing \(N\), since each fragment is tested only once. In practical terms, this means that for small \(f\) (small number of soluble fragments) it is worth increasing the number of fragments per domain \((N)\). On the other hand, when a large number of soluble fragments are available, only a few fragments should be tested before moving to a new domain.

**Intuitive Derivation of** \(P_N\) **for the single copy per fragment case.**

The calculation of \(P_N\) is equivalent to the classical problem of computing the probability of drawing a black marble from a finite set of black and white marbles. It is understandable by considering a probability tree, representing the random generation of a soluble (S) or insoluble (I) fragment. The outcome of \(N\) trials is denoted by a string of Is and Ss, such as SSI. For example, the possible outcomes for \(N = 2\) are SS, SI, IS, and II. The expressions for \(A\), the probability of obtaining one or more soluble fragments per domain, are described as follows. For \(N = 1\),

\[ P_N = P_N(S) = \frac{f}{F}. \]

For \(N = 2\),

\[ P_N = P_N(SS) + P_N(SI) + P_N(IS) \]
\[ = \frac{f}{F} + \frac{F-f}{F} \cdot \frac{f}{F-1}, \]

and for \(N = 3\),

\[ P_N = P_N(SSS) + P_N(SSI) + P_N(SIS) + P_N(SII) + P_N(ISS) + P_N(ISI) + P_N(IIS). \]
\[
= \frac{f}{F} + \frac{F-f}{F} \cdot \frac{f}{F-1} + \frac{F-f}{F} \cdot \frac{F-f-1}{F-1} \cdot \frac{f}{F-2},
\]
where the terms are classified by the position of the first appearance of S. For example, the first term \(f/F\) is derived from \(P_N\) (SSS), \(P_N\) (SSI), \(P_N\) (SIS), and \(P_N\) (SII), in which S appears in the first position; the second term is derived from \(P_N\) (ISS) and \(P_N\) (ISI), in which S is at the second position; the third term is derived from \(P_N\) (IIS), in which S is at the third position. \(N\) determines the size of a denominator, and it is expressed as \(F(F-1)(F-2)\cdots(F-N+1)\). Finally, we obtain:

\[
P_N = \frac{f}{F} \cdot \frac{F-f}{F} \cdot \frac{f}{F-1} \cdots \frac{F-f-N+2}{F-N+2} \cdot \frac{f}{F-N+1} = \sum_{i=0}^{N-1} \frac{(F-f)f(F-i)!}{i!(F-f-i)(F-i)},
\]

where \(i\) represents the \((i+1)\)-th generation of a fragment. The condition \(F-f-i \geq 0\) is required for each term, which is limited by the number of insoluble fragments, \(F-f\), because the terms, except for the first term, suppose the existence of more than one insoluble fragment. If \(F-f-i \geq 0\) does not hold for a term, then it is 0.

**Direct derivation of \(P_N\) for the single copy per fragment case using the hypergeometric distribution**

The hypergeometric distribution is defined as

\[
\binom{s}{i} \binom{F-s}{F-i} \binom{N-s}{i} \binom{N-s}{F-i},
\]

where \(s\) is the number of successes in a sequence of \(N\) draws, \(F\) is the total number, and \(f\) is the number of successes in the total. It represents a probability of the occurrence of \(s\) successes in a sequence of \(N\) draws without replacement. In our situation for the single copy per fragment case, we need \(P_N^C\) (see Eq. 7 in the main text) and thus \(s\) is substituted by 0 in A7. It becomes,
\[ P_N = 1 - P_N^C = 1 - \frac{F-f}{F} \frac{C_N}{C_N} = 1 - \frac{(F-f)(F-N)!}{(F-f-N)!F!} \]  

(A8)

with, \( P_N^C = \frac{f}{F} C_N \times \frac{C_{N-f}}{C_N} = \frac{f}{F} C_N \times \frac{C_{N-f}}{C_N} = \frac{F-f}{F} \frac{C_N}{C_N} \).

Eq. A8 is equivalent to A6 as demonstrated below: \( P_N \) is decomposed as the sum of \( i \)-th series. An \( i \)-th term represents the probability of the occurrence that first \((i-1)\) draws are all failures, an \( i \)-th draw is a success, and the rest of \((N-i)\) draws are unspecified.

Equation A8 is proved as follows:

\[
\sum_{i=0}^{N-1} \frac{(F-f)(F-i) f}{F! (F-f-i)! (F-i)!} = \frac{(F-f) f}{F!} \sum_{i=0}^{N-1} \frac{(F-i) (F-i-1)!}{F! (F-i-1)! (F-i-1)!} = \frac{(F-f) f}{F!} \sum_{i=0}^{N-1} \frac{(F-i-1) (f-1)!}{F! (f-1)! (f-1)!} = \frac{(F-f) f}{F!} \sum_{i=0}^{N-1} \frac{(F-i-1) (f-1)!}{F! (f-i-1) (f-1)! (f-1)!} = \frac{1}{F} C_f \sum_{i=0}^{N-1} F^{-i} C_{f-i}
\]

Using A9 described below, we can obtain

\[
\frac{1}{F} C_f \left( F^{-i} C_{f-i} + F^{-i} C_{f-i} + F^{-i} C_{f-i} + \cdots + F^{-i} C_{f-i} \right) = \frac{1}{F} C_f \left( F^{-i} C_{f-i} \right)
\]

and therefore

\[
\sum_{i=0}^{N-1} \frac{(F-f) f}{F! (F-f-i)! (F-i)!} = \frac{1}{F} C_f \left( F^{-i} C_{f-i} \right) = F^{-i} C_{f-i} = F^{-i} C_{f-i} = F^{-i} C_{f-i} = \frac{1}{F} C_f \left( F^{-i} C_{f-i} \right)
\]

\[
= 1 - \frac{(F-N)!}{(F-N)!} \frac{(F-f)!}{F!} = 1 - \frac{(F-f)!}{(F-f-N)! N!} \frac{(F-N)!}{F!} = 1 - \frac{F-f}{F} \frac{C_N}{C_N}.
\]

(Q.E.D.)

\[
f C_{f-i} = F^{-i} C_{f-i} + F^{-i} C_{f-i} + F^{-i} C_{f-i} + \cdots + F^{-i} C_{f-i}
\]

(A9)

is obtained by repeatedly applying a standard formula,

\[
f C_{f-i} = F^{-i} C_{f-i} + F^{-i} C_{f-i},
\]
to $\phi C_f$ as follows:

$$\phi C_f = \phi_{-1} C_f + \phi_{-1} C_{f-1}$$
$$= \phi_{-2} C_f + \phi_{-2} C_{f-1} + \phi_{-1} C_{f-1}$$
$$= \phi_{-3} C_f + \phi_{-3} C_{f-1} + \phi_{-2} C_{f-1} + \phi_{-1} C_{f-1}$$
$$= \phi_{-4} C_f + \phi_{-4} C_{f-1} + \phi_{-3} C_{f-1} + \phi_{-2} C_{f-1} + \phi_{-1} C_{f-1}$$

$$= \phi_{-N} C_f + \phi_{-N} C_{f-1} + \phi_{-N+1} C_{f-1} + \cdots + \phi_{-2} C_{f-1} + \phi_{-1} C_{f-1}.$$ 

There are $(N+1)$ terms in the right hand side of the last expression and transposing the first term to the left hand side is $A9$. 