# Supporting information:

## Universal encoding of next generation DNA-encoded chemical libraries

Louise Plais,[a] Alice Lessing,[a] Michelle Keller,[a] Adriano Martinelli,[a] Sebastian Oehler,[a] Gabriele Bassi,[b] Dario Neri*[b] and Jörg Scheuermann*[a]

### Table of Contents



[a] Department of Chemistry and Applied Biosciences
Swiss Federal Institute of Technology (ETH Zürich)
Vladimir-Prelog-Weg 4
CH-8093 Zürich (Switzerland)

[b] Philochem AG
Libernstrasse 3
CH-8112 Otelfingen (Switzerland)

## 1. List of abbreviations:

AGP: alpha glycoprotein 1
ATP: adenosine triphosphate
CAII: carbonic anhydrase II
DEL: DNA-encoded chemical library
DNA: deoxyribonucleic acid
dNTP: deoxyribonucleosides triphosphate
ESAC: encoded self-assembling chemical library
HPLC: high performance liquid chromatography
JP: junction primer
LED: large encoding design
LNA: locked nucleic acid
PBS: phosphate buffered saline
PCR: polymerase chain reaction
TEAA: triethylamine and acetic acid solution
T3P: terminal 3'-primer

## 2. Materials and general methods

### a. General

All DNA material was purchased from LGC Biosearch Technologies (Risskov, Denmark), purified by ethanol precipitation, analyzed by LC-MS and quantified with a Nanodrop 2000c Spectrophotometer. T4 DNA polymerase, T4 DNA ligase, ATP, and dNTPs were purchased from New England BioLabs (Ipswich, Massachusetts, USA). Alpha-glycoprotein 1 was purchased from Athens Research (Athens, Georgia, USA) and biotinylated in-house.[1] *N*-[1,1'-biphenyl]-2-yl-*N*-[(4-methylphenyl-9-sulfonyl]-glycine and 5-[2-(trifluoromethyl)methyl)-phenyl]-2-furanpropanoic acid were purchased from Merck (Darmstadt, Germany).

### b. Electrophoresis analyses

DNA was analyzed and purified by 2 to 3% agarose gel, stained with GelRed® Nucleic Acid Stain. The gels were run in TBE buffer, at 110V and 80mA for 30-50 minutes.

SDS-PAGE was performed with pre-casted TBE-Urea gels (15%). The gels were run in TBE buffer, at 200V and 110mA for 60 minutes.

### c. LC-MS analyses

ESI-ToF-MS were performed on a Waters Xevo G2XS Qtof instrument coupled to a Waters Acquity UPLC H-Class System using an Acquity UPLC® Oligonucleotide BEH C18 column, 130Å, 1.7µm, 2.1 x 50mm (Waters) column at 60°C. Hexafluoroisopropanol (0.4M) and triethylamine (0.005M) in water (eluent A) and methanol (eluent B) were used as mobile phase at a flow rate of 0.5 ml/min. Gradient was programmed as follows: after 0.5 min isocratic with 95% eluent A, linear change from 95% eluent A to 50% eluent A in 6.5 min, stepwise back to 100% eluent A in 0.1min and stay for 1min.

### d. Ethanol precipitation

To an aqueous solution of DNA, 0.2 volume of NaCl (5M) and 2.5 volumes of absolute ethanol were added. The solution was kept at -20°C for at least 3hours and centrifuged at 208000 RCF for 30min at 4°C. The supernatant was removed and the sample was dried by a Speedvacuum instrument.

### e. HPLC purifications

Preparative reversed-phase high-pressure liquid chromatography (HPLC) was performed on a Waters Alliance HT RP-HPLC with PDA UV detector, using a X-Terra Shield 5µm, RP 150 × 10 mm C18 column with a gradient of eluent C (TEAA 100 mM) and eluent D (TEAA 100 mM in 80% ACN) at 65°C. Gradient was programmed as follows: after 5min isocratic with 95% eluent C, there was a linear change of 95% eluent C to 82% eluent C in 14min, then again a linear change to 80% eluent D in 5min, and finally a linear change to 100% eluent D for 2min. Then it stayed at 100% eluent D for 1min and stepwise went back to 90% eluent C for 1min.

### f. Splint ligations

The DNA strand (1eq) and the coding DNA (1.1eq) were annealed with a 20-bases long DNA adaptor (1.2eq) by rising temperature 70°C for 10min and then cooling down to 20°C. T4 DNA ligase (400U) was added and the reaction was kept at 16°C for at least one hour. Enzyme was degraded by warming up again to 70°C for 10min.

### g. Formation of amplifiable DNA

Following a previously reported protocol,[2] 0.04nmol of each sub-library member with 0.1nmol of junction primer and terminal primer were mixed together in a total volume of 15µL water with 5µL of T4 DNA ligase buffer and annealed by rising the temperature to 95°C for 2min. The mixture was allowed to cool down to room temperature in one hour. Then, DNA polymerase (10U) and dNTPs (50nmol) were added and the mixture was kept at 16°C for 30min. Finally, T4 DNA ligase (400U) and ATP (30nmol) were added and kept at 16°C for 1hour. Enzymes were degraded at 60°C for 10min.

### h. Coupling of carboxylic acids on DNA

For each compound, an activation solution was made with 90µL DMSO, 42µL carboxylic acid (0.2M in DMSO), 42µL EDC (0.1M in DMSO) and S-NHS (0.333M in DMSO:water, 2:1). A DNA solution was prepared with 10µL universal oligonucleotide (1mM in water, 10nmol scale), 40µL water and 50µL TEA.HCl buffer (pH=10, 50mM). The activation solution was shaken at 37°C for 30min and mixed with the DNA solution. The mixture was subsequently shaken overnight at 37°C and ethanol precipitated.

### 3. DNA sequences:

**Supplementary Table S1:** sequences of the oligonucleotides used in this study.

| Study on one set of codes | |
|---|---|
| Universal oligonucleotide 5'→3' | 5' Amino C6-GGAGCTTCTGAATTCTGTGTGCTG-dSpacer-dSpacer-dSpacer-dSpacer-CTGGTCACTC 3' |
| Universal oligonucleotide 3'→5' | 5' Phos-AGTCACCTCAdSpacer-dSpacer-dSpacer-dSpacer-CAGCACACAGAATTCAGAAGCTCC-C6 Amino 3' |
| Code1_5'→3' | 5' Phos-GTAGTCTCTCAGTCTGCTGTCGTACG 3' |
| Code2_5'→3' | 5' Phos-TTGCTCACACATCGATTGTCAACTCGGTCCTG 3' |
| Code1_3'→5' | 5' Phos-GAAGGGCTACAGTTCGTTCGCTCGCT 3' |
| Code2_3'→5' | 5' CTTGACGATCGATGGAGTGATTTGAGTGAGTG 3' |
| Adaptor universal oligonucleotide-code1_5'→3' | 5' GAGAGACTACGAGTGACCAG 3' |
| Adaptor code1-code2_3'→5' | 5' GTGTGAGCAACGTACGACAG 3' |
| Adaptor universal oligonucleotide-code1_5'→3' | 5' TGAGGTGACTAGCGAGCGAA 3' |
| Adaptor code1-code2_3'→5' | 5' GTAGCCCTTCCACTCACTCA 3' |
| Junction primer (JP) | 5' Phos-GAGAGACTACGAGTGACCAGTTTGAGGTGACTAGCGAGCGAA 3' |
| Terminal 3'-primer (T3P) | 5' ATCAGTTCATGGGTATGCGACAGGACCGAGTTGAC 3' |
| Ligation efficiency assessment: second set of codes | |
| Code1'_5'→3' | 5' Phos-GTAGTCTCTCGAGAAGCTGTCGTACG 3' |
| Code2'_5'→3' | 5' Phos-TTGCTCACACGTGGGACGTCAACTCGGTCCTG 3' |
| Code1'_3'→5' | 5' Phos-GAAGGGCTACGAGAAGTTCGCTCGCT 3' |
| Code2'_3'→5' | 5' CTTGACGATCGATGGGTGGGACTGAGTGAGTG 3' |
| Model selection experiment: set of codes | |
| AGP-code1_5'→3' | 5' Phos-GTAGTCTCTCTGGCGTCTGTCGTACG 3' |
| CAII-code1_5'→3' | 5' Phos-GTAGTCTCTCGAGAAGCTGTCGTACG 3' |
| AcOH-code1_5'→3' | 5' Phos-GTAGTCTCTCTGAGAGCTGTCGTACG 3' |
| AGP-code1_3'→5' | 5' Phos-GAAGGGCTACTGGCGTTTCGCTCGCT 3' |
| CAII-code1_3'→5' | 5' Phos-GAAGGGCTACGAGAAGTTCGCTCGCT 3' |
| AcOH-code1_3'→5' | 5' Phos-GAAGGGCTACTGAGAGTTCGCTCGCT 3' |
| AGP-code2_5'→3' | 5' Phos-TTGCTCACACTGGCGTAGTCAACTCGGTCCTG 3' |
| CAII-code2_5'→3' | 5' Phos-TTGCTCACACTATCCCCGTCAACTCGGTCCTG 3' |
| AcOH-code2_5'→3' | 5' Phos-TTGCTCACACCCTTCCCGTCAACTCGGTCCTG 3' |
| AGP-code2_3'→5' | 5' ATCCTTGACGATCGATGGCCTTCCCTGAGTGAGTG 3' |
| CAII-code2_3'→5' | 5' ATCCTTGACGATCGATGGCCTTCCCTGAGTGAGTG 3' |
| AcOH-code2_3'→5' | 5' ATCCTTGACGATCGATGGCCTTCCCTGAGTGAGTG 3' |

## 4. DNA characterization

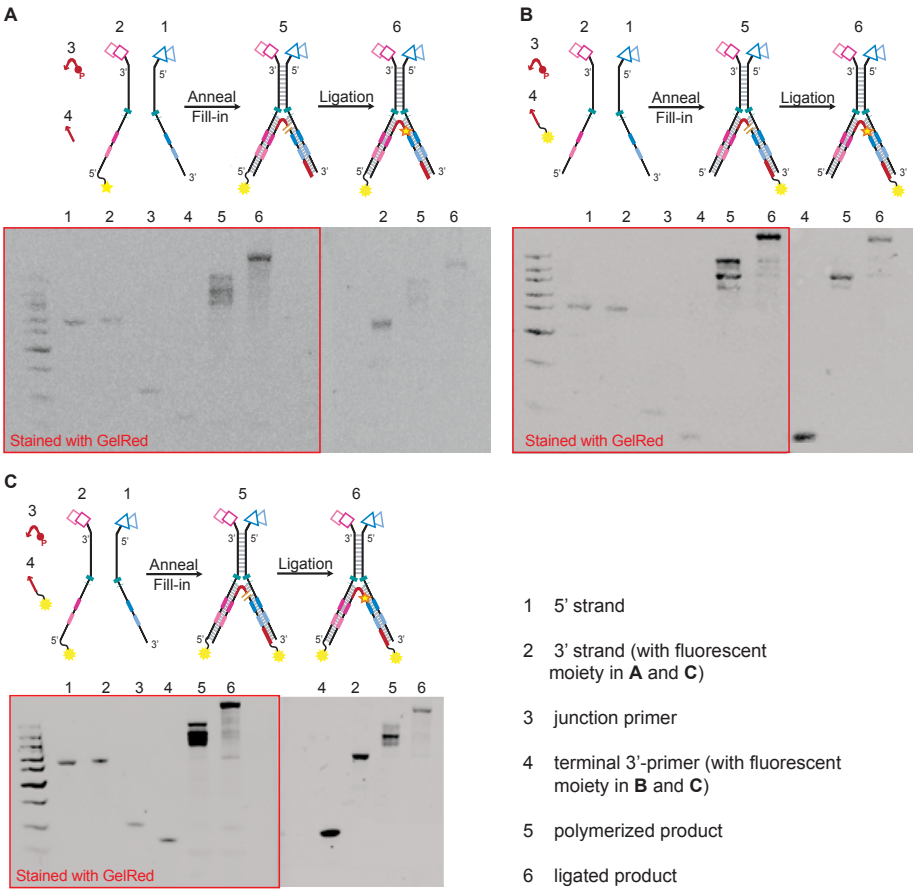
**Supplementary Table S2:** oligonucleotides description for Figure 2.

| DNA | N° of nucleotides |
|---|---|
| I (Figure 2A) | 64 |
| II (Figure 2A) | 96 |
| III (Figure 2A) | 64 |
| IV (Figure 2A) | 96 |
| V (Figure 2A) | 38 |
| VI (Figure 2A) | 26 |
| VII (Figure 2A) | 20 |
| VIII (Figure 2A) | 32 |
| IX (Figure 2A) | 20 |
| XIII (Figure 2C) | 124 |

**Supplementary Table S3:** oligonucleotides description and MS analysis for Figures 2B and C.

| DNA | Calculated mass (g/mol) | Observed mass (g/mol) |
|---|---|---|
| X (Figure 2B, 1) JP (red) | 13181 | 13181 |
| X (Figure 2B, 1) T3P (orange) | 10836 | 5418($M+2H^+/2$), 10836 |
| X (Figure 2B, 1) (II) (blue) | 29140 | 29142 |
| X (Figure 2B, 1) (IV) (purple) | 29325 | 29327 |
| XI (Figure 2B, 2) extended-(II) (blue) | 35263 | 35266 |
| XI (Figure 2B, 2) (IV) (purple) | 29325 | 29327 |
| XI (Figure 2B, 2) extended-T3P (dark green) | 21093 | 21095 |
| XI (Figure 2B, 2) extended-JP | 27779 | 27781 |
| XII (Figure 2B, 3) extended-(II) (blue) | 35263 | 17634($M+2H^+/2$), 35268 |
| XII (Figure 2B, 3) (IV) (purple) | 29325 | 29329 |
| XII (Figure 2B, 3) amplifiable DNA (orange) | 48855 | 48861 |
| A (Figure 3A) (II) | 29206 | 14606($M+2H^+/2$), 29212 |
| A (Figure 3A) (IV) | 23019 | 23023 |
| A (Figure 3A) extended-T3P | 17899 | 17903 |
| A (Figure 3A) extended-JP | 18627 | 18631 |
| A (Figure 3A) impurity (orange) | 21052 | 21057 |
| A (Figure 3A) amplifiable DNA | 36540 | 36515 |
| B (Figure 3A) (II) | 29566 | 29573 |
| B (Figure 3A) (IV) | 23019 | 23025 |
| B (Figure 3A) extended-T3P | 17899 | 17903 |
| B (Figure 3A) extended-JP | 19221 | 19226 |
| B (Figure 3A) amplifiable DNA | 37102 | 37111 |
| C (Figure 3A) (II) | 35263 | 17634($M+2H^+/2$), 35268 |
| C (Figure 3A) (IV) | 29325 | 29329 |
| C (Figure 3A) amplifiable DNA | 48855 | 48861 |

## 5. Ligation efficiency



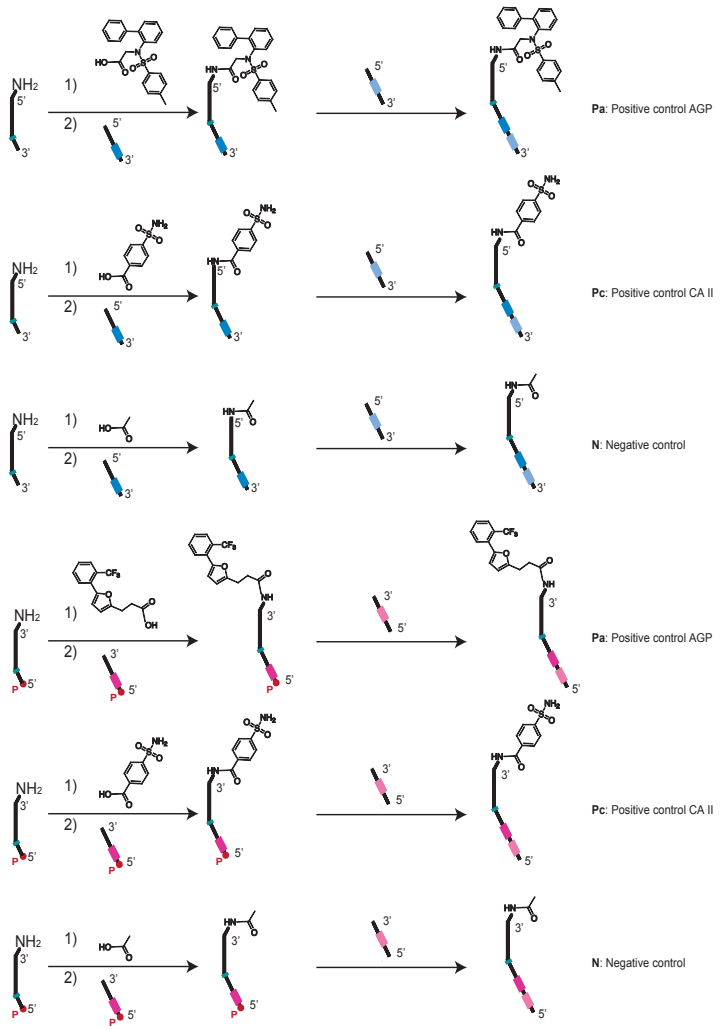

**Supplementary Figure S1.** Ligation step – reliability assessment. Schematic representation of the oligonucleotide sets used and SDS-PAGE gels. Gels were run, cut and the left parts were stained with GelRed® to reveal the non-fluorescent oligonucleotides. Yellow stars represent the fluorescent tag FAM. **A.** Strand **2** tagged with FAM. The fluorescent band shifts in SDS-PAGE gel demonstrating full conversion to the desired product after fill-in and ligation. **B.** Terminal 3'-primer **4** tagged with FAM. The fluorescent band shifted in SDS-PAGE gel demonstrating full conversion to the desired product after fill-in and ligation. **C.** Terminal 3'-primer **4** and of strand **2** tagged with FAM. The fluorescent bands shifted in SDS-PAGE gel demonstrating full conversion to the desired product after fill-in and ligation.

## 6. Selection experiments

### a. Preparation of model libraries

Each DNA strand composing the sub-libraries was synthetized individually. Each sub-library was composed of three distinct strands. *N*-[1,1'-biphenyl]-2-yl-*N*-[(4-methylphenyl-9-sulfonyl]-glycine (CAS: 94870-32-3) or 4-sulfamoylbenzoic acid (SABA) (CAS: 138-41-0) or acetic acid (CAS: 64-19-7) were coupled as described in section 2.h. on universal oligonucleotide 5'→3' and HPLC purified. Corresponding codes 1 and codes 2, 5'→3', were subsequently added by splint ligation and HPLC purified. In the same manner, 5-[2-(trifluoromethyl)methyl)-phenyl]-2-furanpropanoic acid (CAS: 853310-21-1) or 4-sulfamoylbenzoic acid (SABA) or acetic acid were coupled on universal oligonucleotide 3'→5' and HPLC purified. Corresponding control codes 1 and codes 2, 3'→5', were subsequently added by splint ligation and HPLC purified. All products were HPLC purified and analyzed by LC-MS.

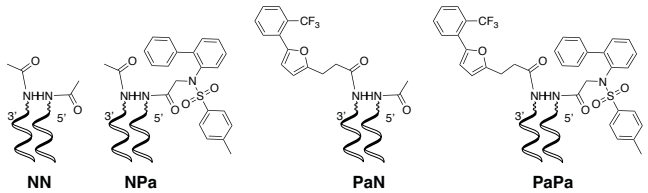
This process was done for LED and ESAC systems (with only one code position for ESAC)[3].




**Supplementary Figure S2**: Schematic representation of the synthesis of each LED library member. The carboxylic acids were coupled on amino-modified DNA strands and redundantly encoded by two ligation steps. P: phosphate.

Model libraries for the first series of selections:

A first set of heteroduplexes collections (model libraries) was formed with the previously modified DNA strands. Positive controls for AGP and negative controls were mixed to obtain the four duplexes described in **Supplementary Figure S3** in a 1:1:1:1 ratio. On a 20pmol scale, the complementary DNA strand were annealed and the amplifiable DNA strands were formed as described above for LED and ESAC. Then libraries were diluted to $10^7$, $10^8$ $10^9$, $10^{10}$ and $10^{12}$ molecules/5µL for the selection experiments.

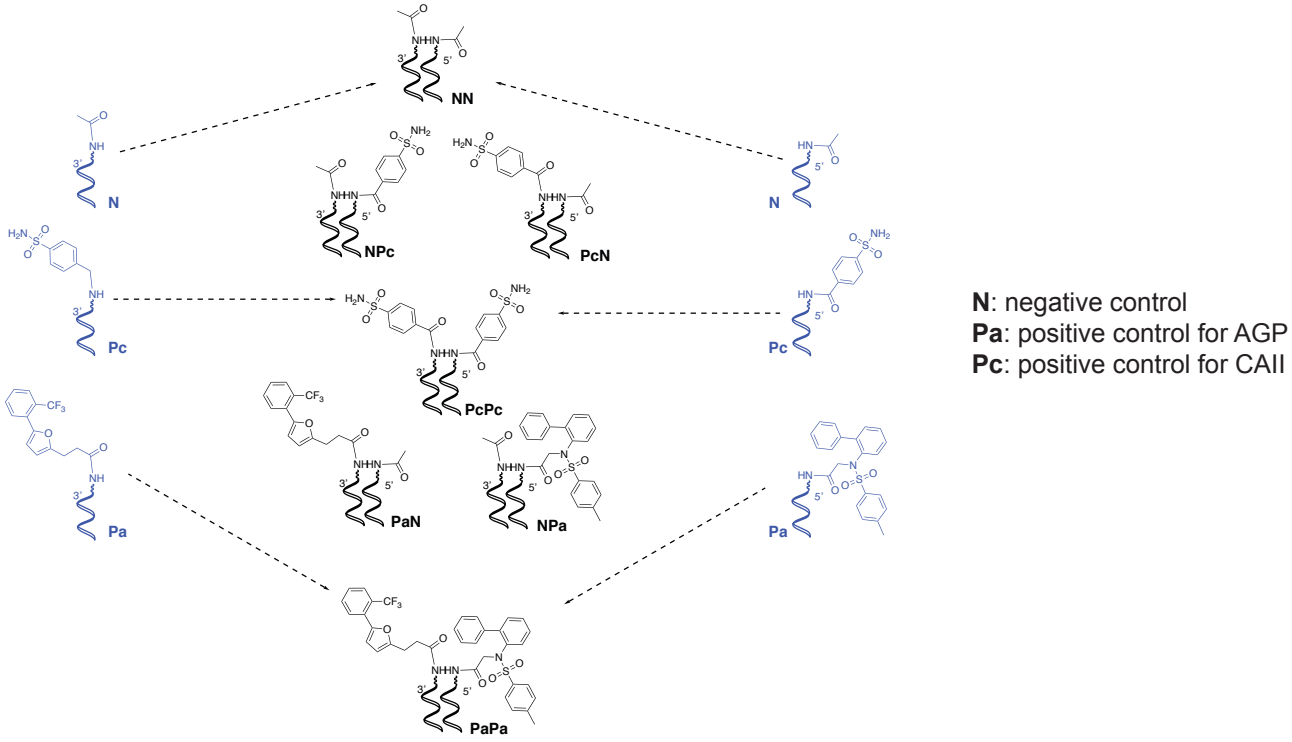



**N**: negative control
**Pa**: positive control for AGP

**Supplementary Figure S3**: Schematic representation of the first mixture of duplexes formed for the first set of model selections. Ratio 1:1:1:1. N=negative control, Pa=positive control for AGP. As a convention, the 3'-strand is mentioned before the 5'-strand.

Model libraries for the second series of selections:

A second set of model libraries was formed with the previously modified DNA strands. Positive controls for AGP and CAII and negative controls were mixed to obtain all combinations of positive and negative controls as depicted in **Supplementary Figure S4**. In total, six model libraries were synthesized with different ratio of positive and negative controls. The theoretical production of each duplex for each model library is displayed in **Supplementary Table S4**, in percentages over the total number of heteroduplexes in a model library.




**Supplementary Figure S4**: schematic representation of the second set of duplexes formed for the second set of model selections. N=negative control, Pa=positive control for AGP, Pc=positive control for CAII. As a convention, the 3' strand is mentioned before the 5'strand. The duplexes are assembled from the sub-library members depicted in blue on the sides. Some library members are represented in black in the middle.

**Supplementary Table S4:** theoretical representations of duplexes in the model libraries used for the second set of model selections. Representations are given in percentages over the total number of heteroduplexes in each model library, kept constant at $10^{13}$. Representations are calculated according to Formula S1.

| Model library | NN | NPa / PaN / NPc / PcN | PaPa / PcPc / PaPc / PcPa | Number of positive control (PaPa or PcPc) in selections | Dilution of positive controls |
|---|---|---|---|---|---|
| **ML1** | ~100 % | 3.16. $10^{-3}$ % | $10^{-7}$ % | $10^{4}$ | 1:$10^{9}$ |
| **ML2** | 99.99 % | 1.00. $10^{-2}$ % | $10^{-6}$ % | $10^{5}$ | 1:$10^{8}$ |
| **ML3** | 99.97 % | 3.16. $10^{-2}$ % | $10^{-5}$ % | $10^{6}$ | 1:$10^{7}$ |
| **ML4** | 99.90 % | 1.00. $10^{-1}$ % | $10^{-4}$ % | $10^{7}$ | 1:$10^{6}$ |
| **ML5** | 99.68 % | 3.16. $10^{-1}$ % | $10^{-3}$ % | $10^{8}$ | 1:$10^{5}$ |

**b. Selection experiments and PCR amplification**

Selections were performed as previously described[1] on a KingFisher Flex magnetic particle processor. Briefly, biotinylated alpha-glycoprotein 1 (AGP) or carbonic anhydrase II (CAII) was incubated with streptavidin magnetic beads. After three washes, bound proteins were incubated with the desired model library and then submitted to five additional washes. Finally, the DNA was eluted in Tris buffer and submitted to two rounds of PCR amplification. The primers for PCR1 bore codes specific for the selection experiment and the primers for PCR2 bore Illumina compatible sequences. Each experiment was done in triplicate. Each model library was selected against uncoated streptavidin magnetic beads for negative control and each model library was also amplified without going through a selection experiment. The last are referred as unselected libraries.

For the first series of selection experiments, different amounts of duplexes from the first set of duplexes were submitted to selections, ranging from $10^{8}$ to $10^{12}$ duplexes with a ratio kept constant at 1:1:1:1.

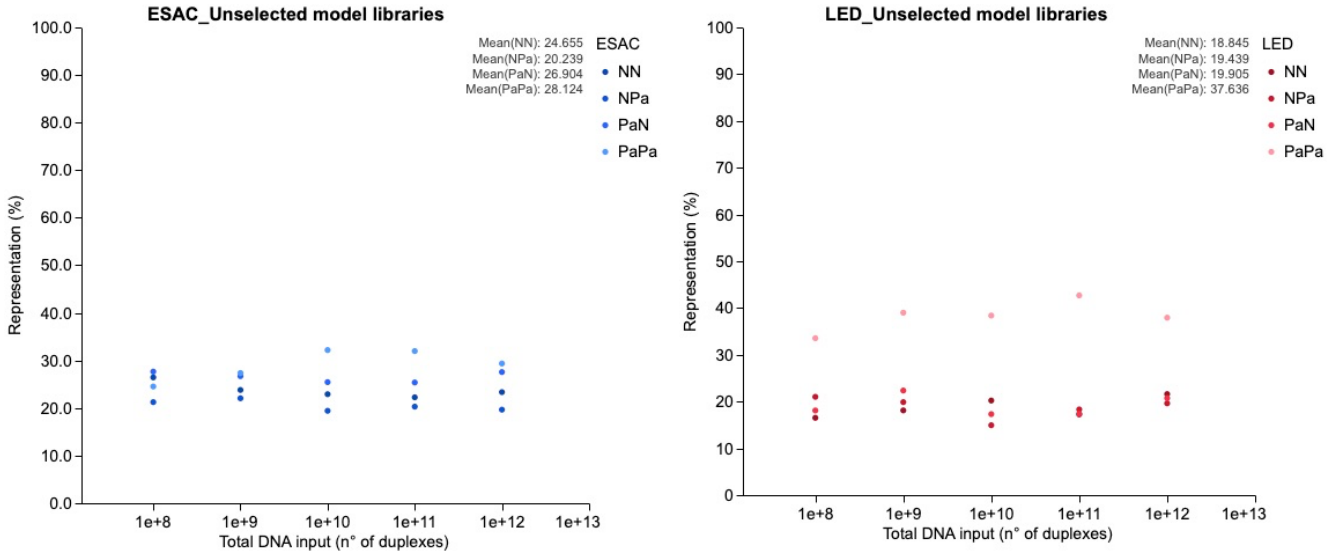
For the second series of selection experiments, a determined number of duplexes from the second set of duplexes was submitted to selections, kept constant at $10^{13}$ duplexes with ratios varying as described in **Supplementary Table S4**. PCR amplification was performed following a previously described protocol.[1] For LED, annealing temperature was lowered to 60°C during the PCR cycles. For the unselected libraries, PCR were started from $10^{8}$ to $10^{12}$ duplexes to reflect on the selection experiments.

**c. Sequencing**

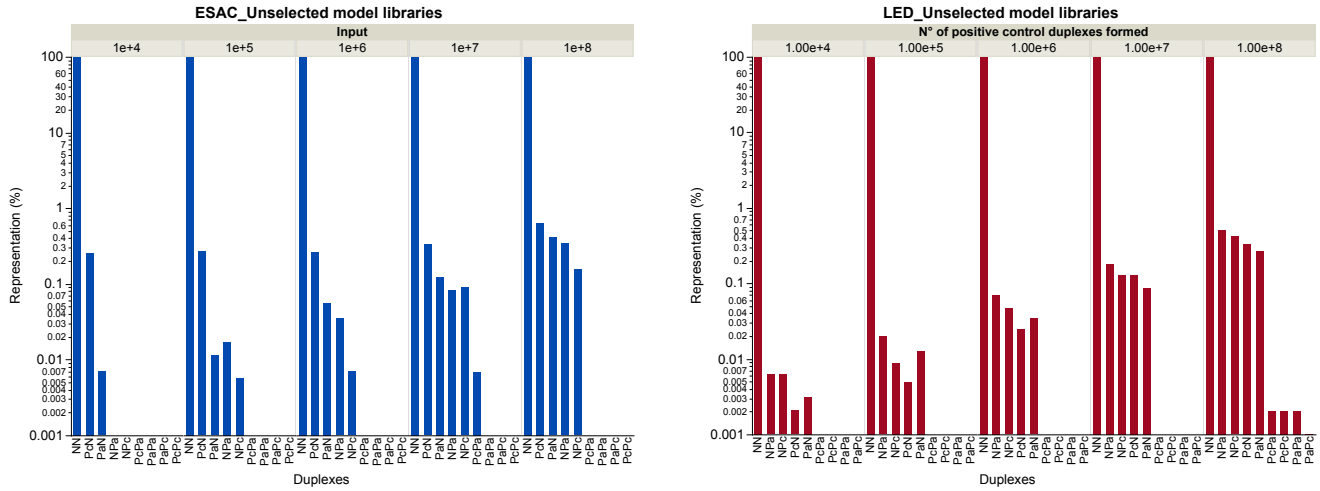
The amplicons were quality-checked with Nanodrop, Qubit and TapeStation instruments. Next Generation Sequencing of the PCR amplicons was carried on a MiSeq instrument. The obtained sequences were analyzed with the help of in-house C++ code. Graphs were plotted with JMP® Pro 15.

**d. Unselected libraries**

First, the unselected libraries were sequenced. The results are displayed in **Supplementary Figure S5** corresponding to the model libraries used in the first set of selections and in **Supplementary Figure S6** for the model libraries used in the second set of selections. In both cases, sequencing data showed the expected representation for each duplex in each model library.



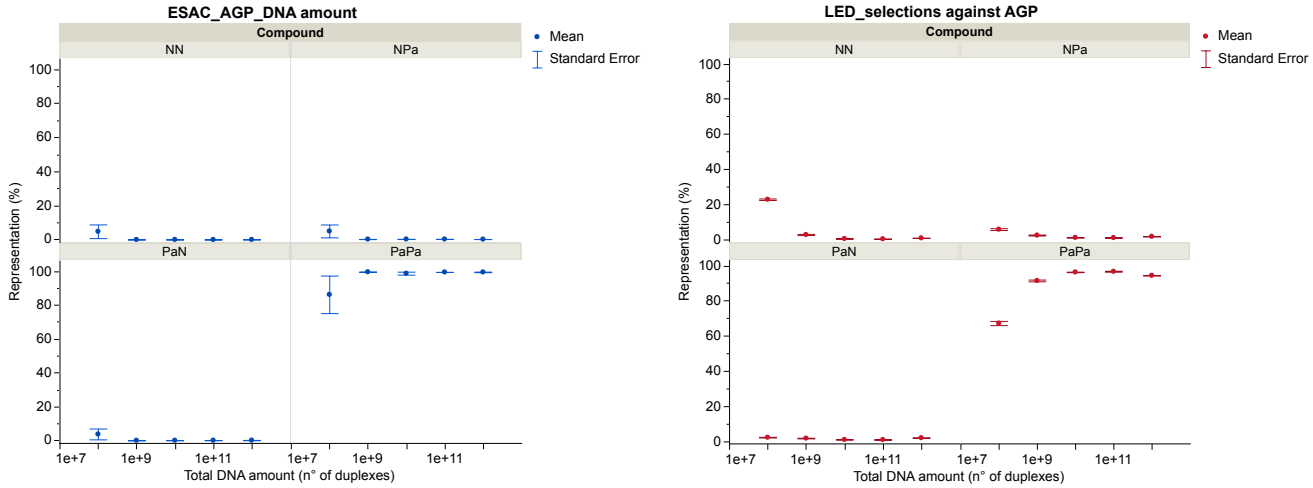

**Supplementary Figure S5:** Unselected model libraries for the first set of selections. The model libraries were PCR amplified without going through selection. The representation of each duplex was calculated following Formula S1 and is displayed as a percentage, depending on the starting total input of duplexes into the PCR. ESAC model libraries are displayed in blue on the left graph and LED model libraries are displayed in red on the right graph.




**Supplementary Figure S6:** Unselected model libraries for the second set of selections. The model libraries were PCR amplified without going through selection. The representation of each duplex was calculated following Formula S1 and is displayed as a percentage in logarithmic scale, depending on the starting input of positive control duplexes. Left: ESAC model libraries are displayed in blue. Right: LED model libraries are displayed in red. The results correspond to the theoretical values displayed in **Table S4**.

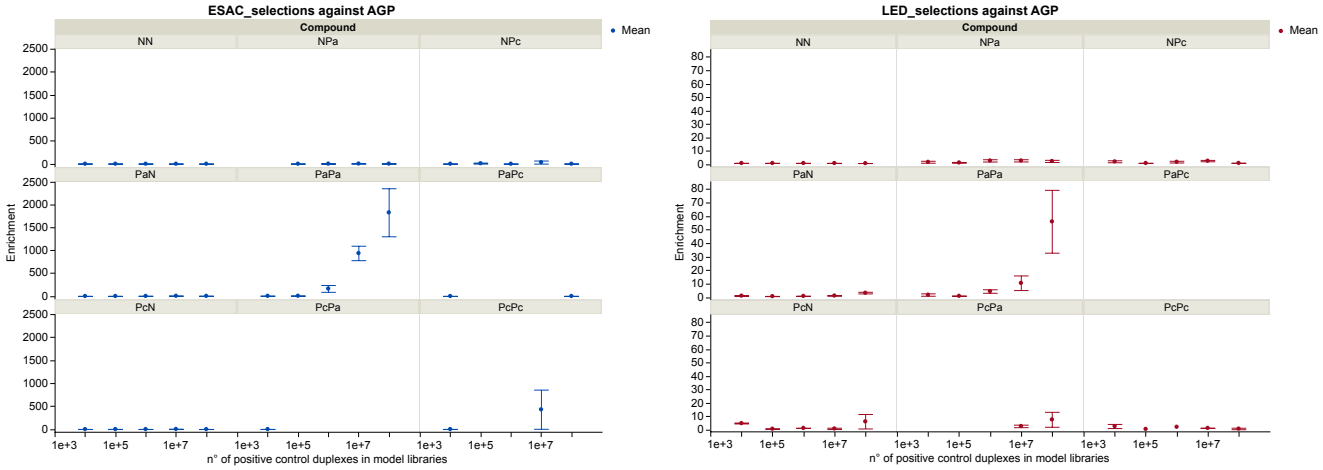
**e. Selection results**

First set of selections

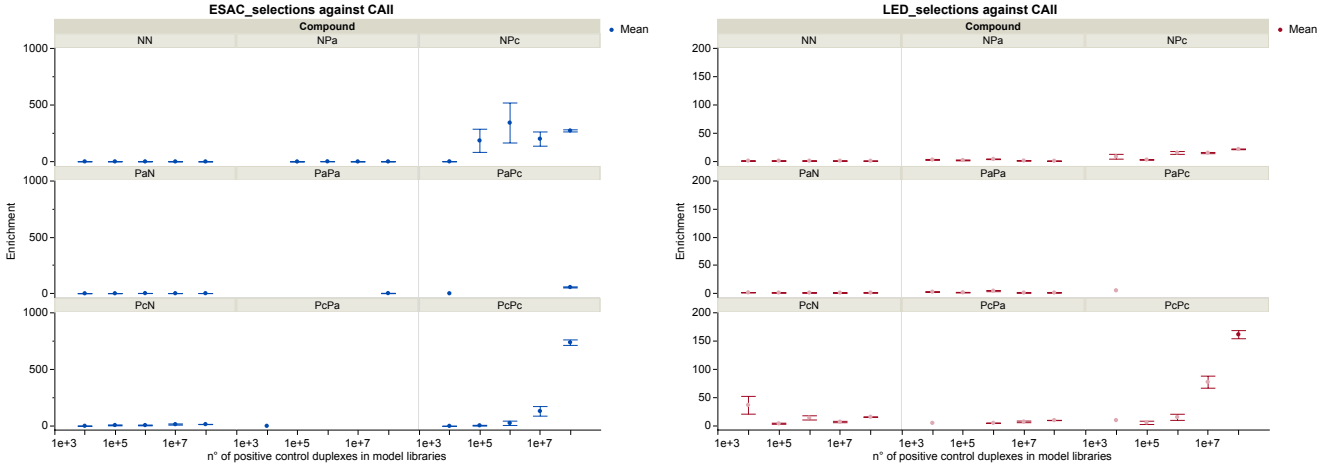



**Supplementary Figure S7:** Selection experiments against AGP – variation of the total DNA amount into the selections. The representation of each duplex was calculated following Formula S1 and is displayed as a percentage, depending on the total amount of DNA duplexes introduced into the selection experiment. Each duplex representation is displayed in a separated window. As expected, only the positive control for AGP (PaPa) was over-represented after the selection experiments. Left: ESAC results are displayed in blue. Right: LED results are displayed in red.

Second set of selections




**Supplementary Figure S8:** Selection experiments against AGP – variation of the positive control dilution into the selections. The enrichment of each duplex was calculated following formula S2 and is displayed as a function of the starting input of positive control duplexes into the selection experiment, while the total DNA input is kept constant at $10^{13}$ duplexes. Each duplex representation is displayed in a separated window. As expected, only the positive control for AGP (PaPa) was significantly enriched after the selection experiments. Left: ESAC results are displayed in blue. Right: LED results are displayed in red.




**Supplementary Figure S9:** Selection experiments against CAII – variation of the positive control dilution into the selections. The enrichment of each duplex was calculated following formula S2 and is displayed as a function of the starting input of positive control duplexes into the selection experiment, while the

total DNA input is kept constant at $10^{13}$ duplexes. Each duplex representation is displayed in a separated window. As expected, only the positive control for CA II (PcPc) was significantly enriched after the selection experiments. Some duplexes displaying one SABA molecules were also enriched, for example NPc. Left: ESAC results are displayed in blue. Right: LED results are displayed in red.

**f. Calculations**

After DNA sequencing and deconvolution, sequence counts were assigned to each duplex for each individual selection experiment. Representation and enrichment were calculated using the following formulae:

**Supplementary Formula S1:**

$$representation\ (\%) = \frac{counts\ for\ duplex\ \boldsymbol{x}\ in\ selection\ against\ \boldsymbol{y}\ protein}{counts\ for\ all\ duplexes\ in\ selection\ against\ \boldsymbol{y}\ protein} \times 100$$

**Supplementary Formula S2:**

$$enrichment = \frac{\frac{counts\ for\ duplex\ \boldsymbol{x}\ in\ selection\ against\ \boldsymbol{y}\ protein}{counts\ for\ all\ duplexes\ in\ selection\ against\ \boldsymbol{y}\ protein}}{\frac{counts\ for\ duplex\ \boldsymbol{x}\ in\ negative\ control\ selection}{counts\ for\ all\ duplexes\ in\ negative\ control\ selection}}$$

**Supplementary Formula S3:**

$$normalized\ enrichment\ (\%) = \frac{enrichment\ for\ duplex\ \boldsymbol{x}\ in\ selection\ against\ \boldsymbol{y}\ protein}{maximum\ enrichment\ in\ selection\ against\ \boldsymbol{y}\ protein} \times 100$$

## 7. References


[1] W. Decurtins, M. Wichert, R. M. Franzini, F. Buller, M. A. Stravs, Y. Zhang, D. Neri, J. Scheuermann, *Nat. Protoc.* **2016**, *11*, 764–780.

[2] Y. Zhou, C. Li, J. Peng, L. Xie, L. Meng, Q. Li, J. Zhang, X. D. Li, X. Li, X. Huang, et al., *J. Am. Chem. Soc.* **2018**, *140*, 15859–15867.

[3] M. Wichert, N. Krall, W. Decurtins, R. M. Franzini, F. Pretto, P. Schneider, D. Neri, J. Scheuermann, *Nat. Chem.* **2015**, *7*, 241–249.